WHAT IS GENOMICS AND WHY SHOULD I CARE?

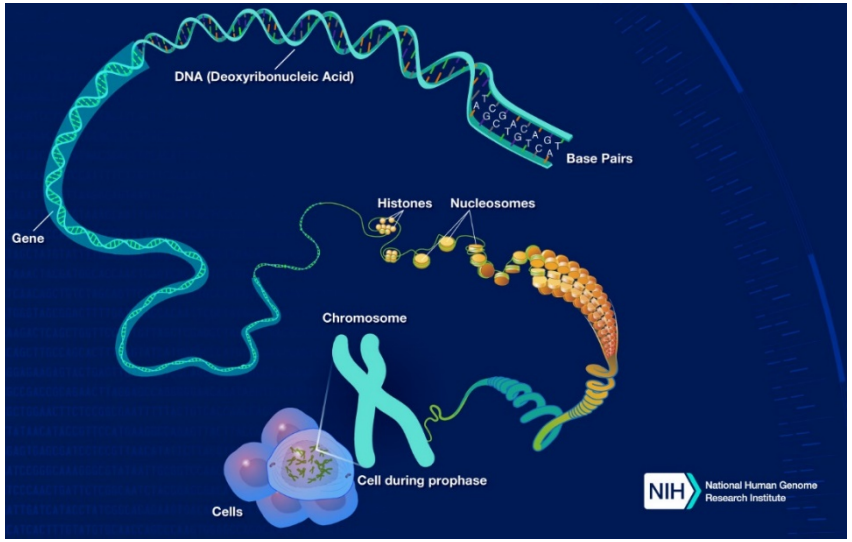
David A Ray
Professor, Dept. of Biological Sciences
Texas Tech University

Genetics vs. Genomics

- The main difference:
  - Genetics typically examines the function and composition of single genes
  - Genomics typically addresses all functional aspects of a genome and their interrelationships

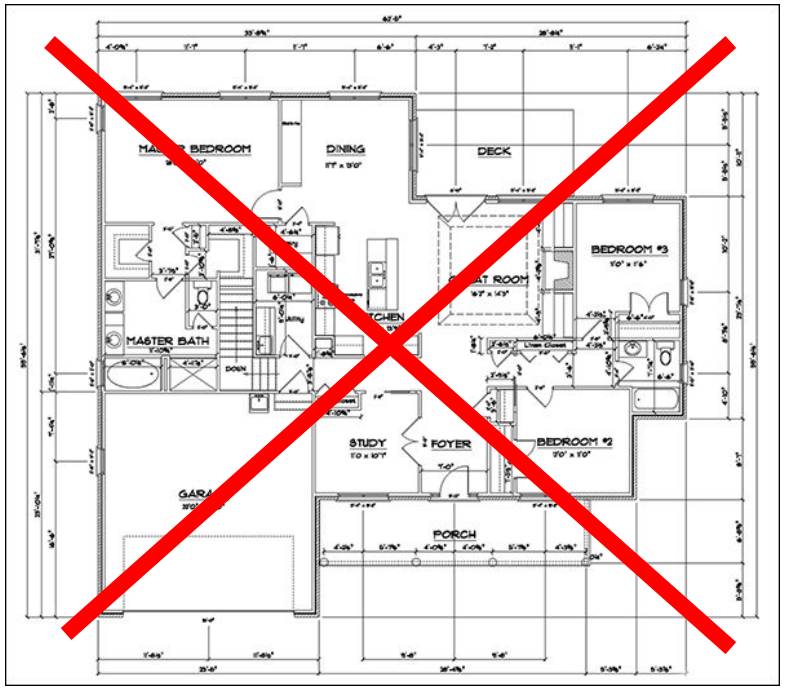
Genome

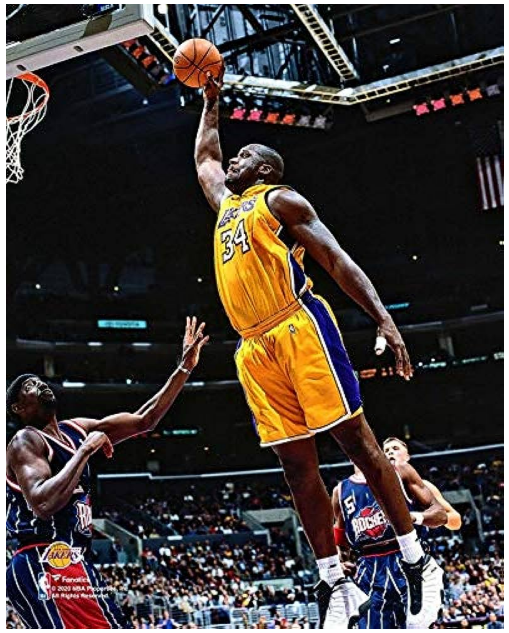
- Generic definition: Minimum nucleic acid complement that defines an organism/organelle/virus
  - Will vary depending on organism, organelle, or virus
  - Mostly DNA but some viruses have RNA-based genomes



Is the genome a blueprint?

- No. A blueprint is a scaled copy, *a genome is a chemical that interacts with the cellular and external environments to direct biochemical processes*
  - The DNA of our genome both facilitates and restricts the growth and function of a cell or multiple cells but does not **dictate** it.
  - Environmental impacts
    - Toxins, nutrition, exposure to disease
  - If someone were to clone you…?





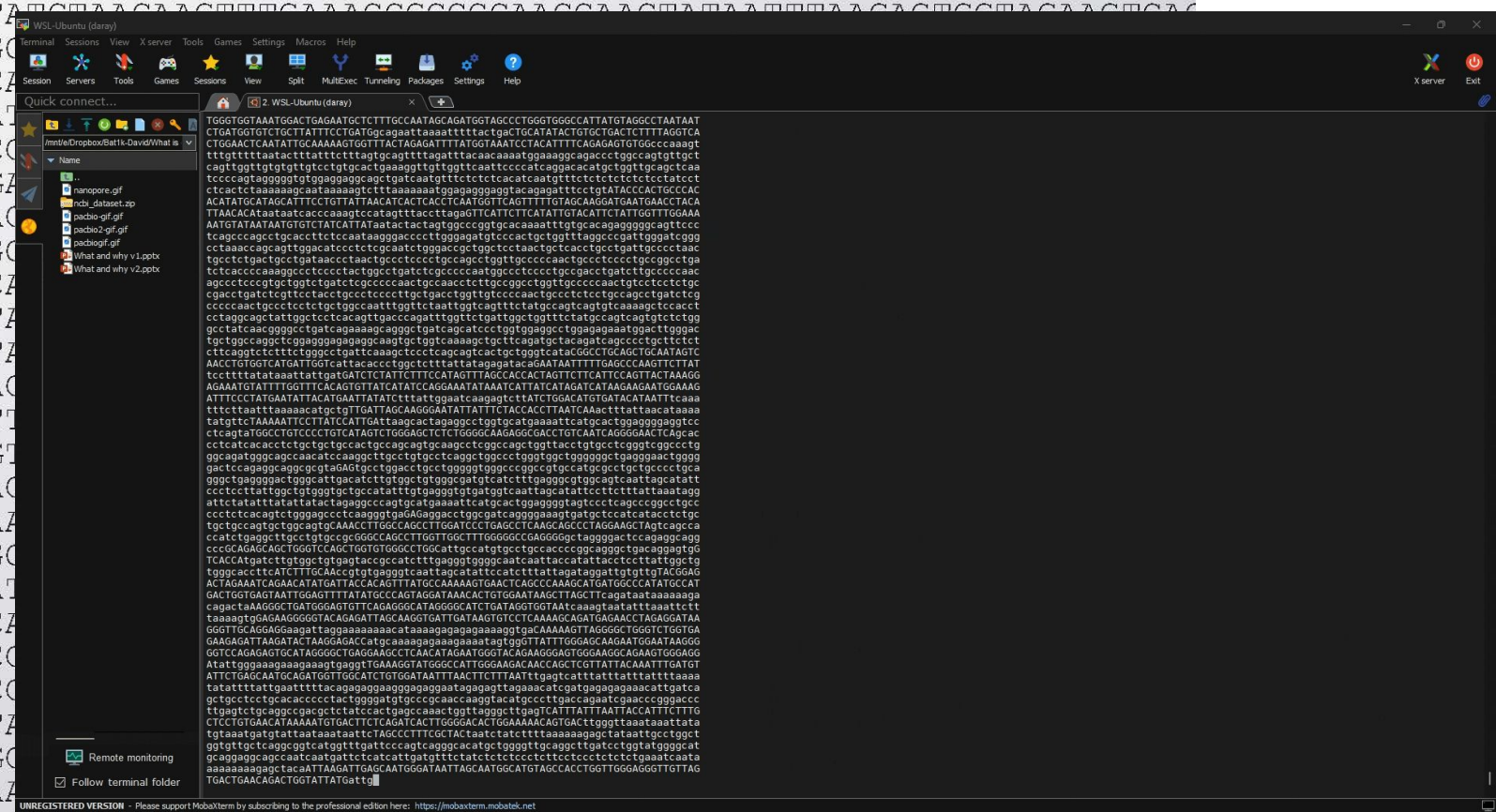
Genomics

- Research in which robotics, automated sequencing, and advanced computational methods are utilized to *rapidly* and *efficiently* characterize genomes and their components
- Genomes vs. genome assemblies
- A **genome** is a physical thing in organisms that's *made up of DNA*
- A **genome assembly** is an approximation of a genome *made up of bits and bytes*
- While genomes can be investigated directly, we often use assemblies as their proxies.

Genomics

- Many different subfields
  - Applied genomics
  - Population genomics
  - Regulatory genomics
  - Structural genomics
  - Bioinformatics
  - Metagenomics
  - Functional genomics
  - Disease genomics
  - Comparative genomics
  - Phylogenomics

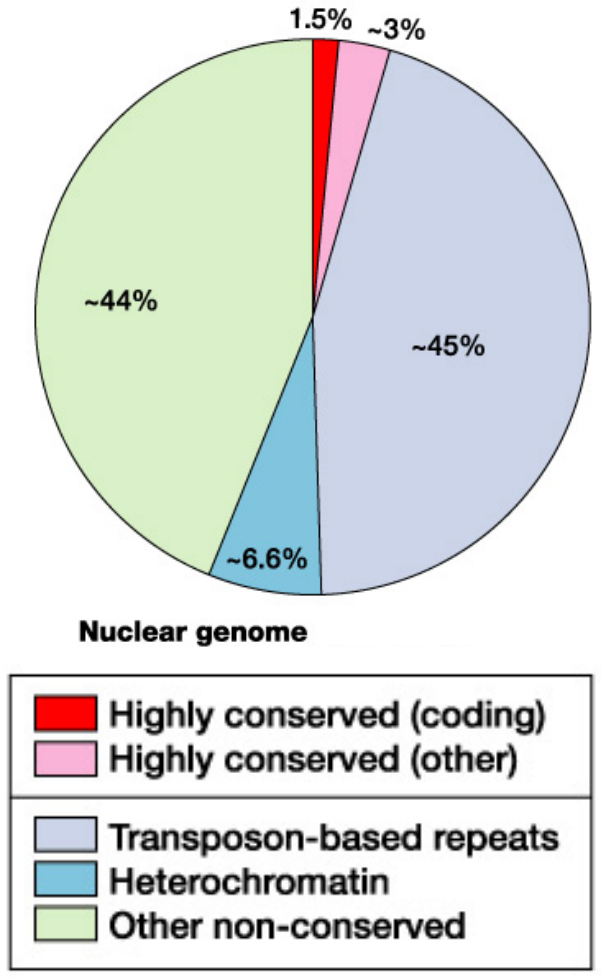
What do genomes look like and how do we make sense of them?



Genome Components

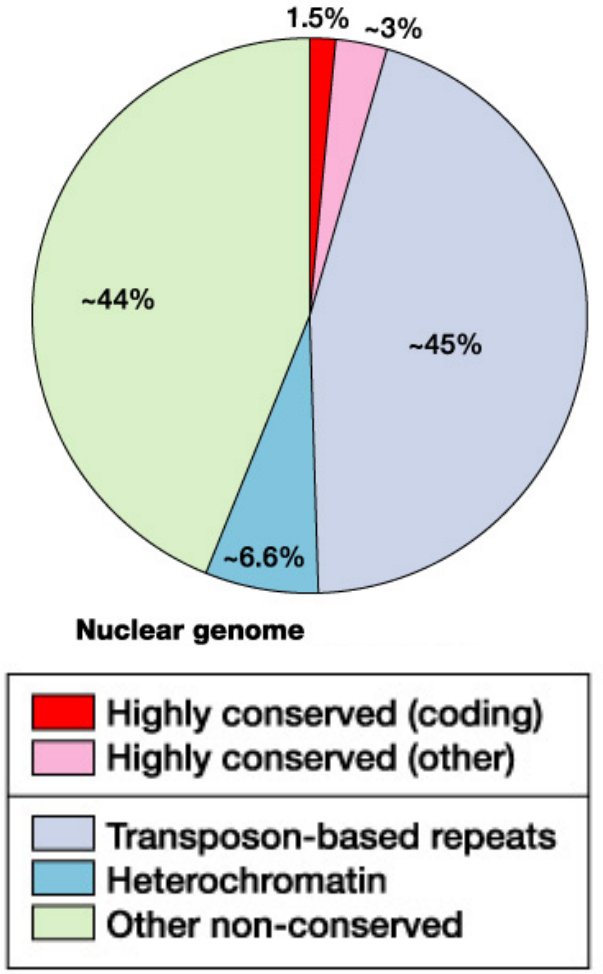
- Be aware
  - that many parts of a genome may belong to several categories
  - that I am primarily talking about complex, eukaryotic genomes
- Non-coding DNA
- Repetitive DNA
- Coding DNA

The human genome as an example



Genes vs. alleles vs. loci

- **Gene**: “Region of DNA that controls a discrete hereditary characteristic, often (but not always) corresponding to a single protein or RNA. This definition includes the entire functional unit, encompassing coding DNA sequences, non-coding regulatory DNA sequences, and introns.”
- **Allele**: “One of a set of alternative forms of a gene.”
- **Locus**: “The position of a gene on a chromosome. Different alleles of the same gene all occupy the same locus.”




1.5%
~3%
~44%
~45%
~6.6%
Nuclear genome
Highly conserved (coding)
Highly conserved (other)
Transposon-based repeats
Heterochromatin
Other non-conserved


Genes
Encode proteins
Regulatory motifs
Control gene expression
1.5%
~3%
~44%
~45%
~6.6%
Nuclear genome
Highly conserved (coding)
Highly conserved (other)
Transposon-based repeats
Heterochromatin
Other non-conserved

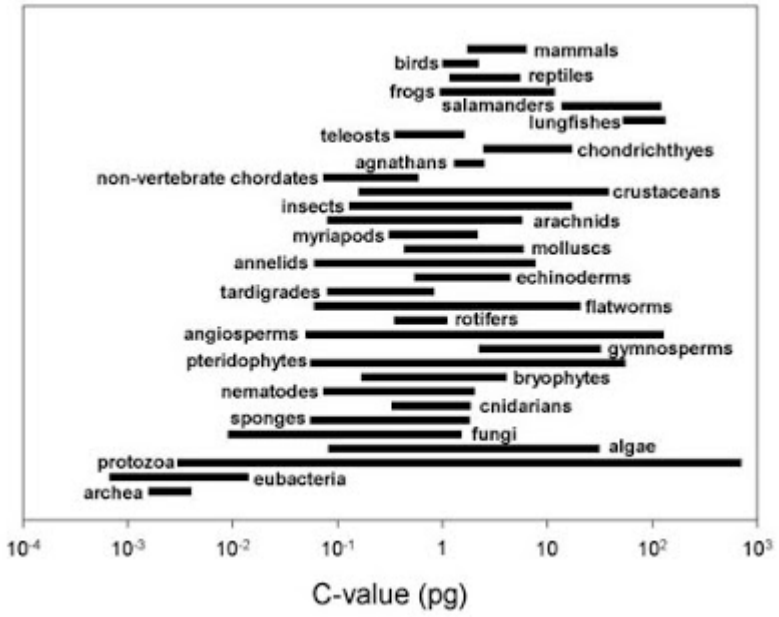
Sizes and Organization of Genomes

- There is no “one” genomic sequence for a species
- The differences in DNA sequence among members of the same species are generally relatively few, but they can be very important.

- **Variant**: any individual variation in the genomic content when compared to the population as a whole. Could be a SNP, indel, CNV, etc.

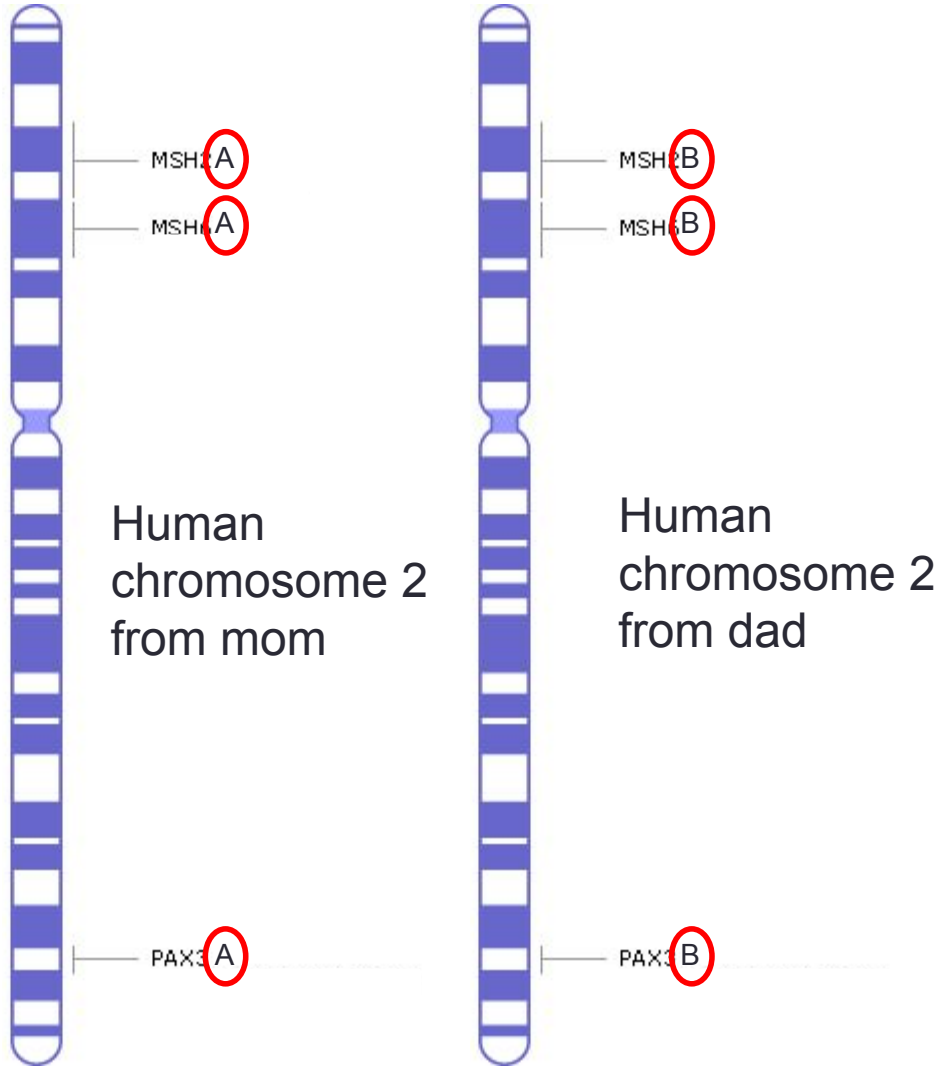
Genome Sizes

- **Genome Size and Organism Complexity**
  - Very rough, imperfect correlation
  - C-value – the characteristic value of haploid DNA content per nucleus; aka how much DNA is there in a genome?
  - *C-value paradox*: the observation that genome size does not correlate with organismal complexity
  - Taft et al. (2007) - complexity can be "broadly defined as the number and different types of cells, and the degree of cellular organization."
  - *The onion test is a simple reality check for anyone who thinks they have come up with a universal function for non-coding DNA. Whatever your proposed function, ask yourself this question: Can I explain why an onion needs about five times more non-coding DNA for this function than a human?*



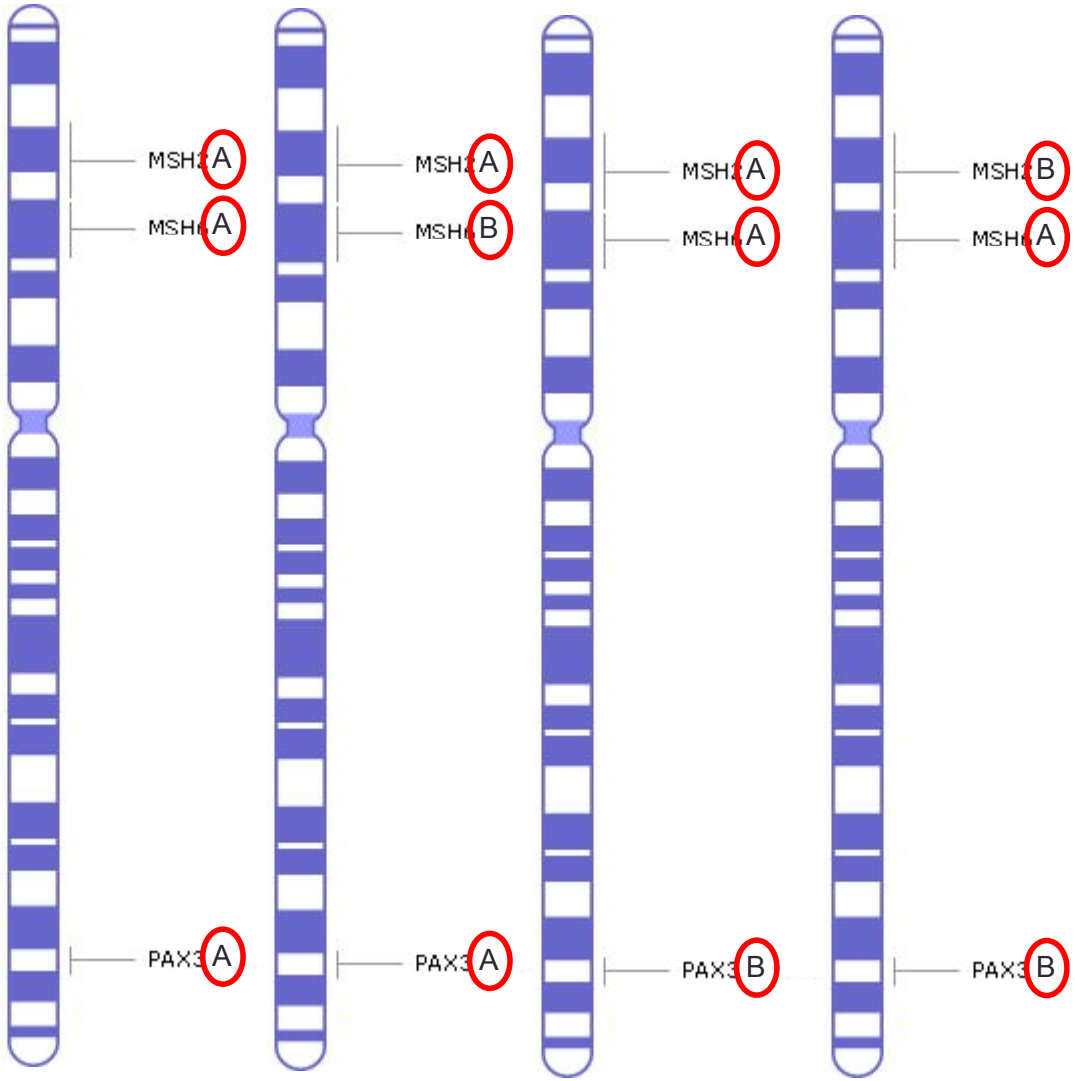
Chromosomes and genome maps

- DNA in most eukarotes is organized into linear chromosomes
  - Having *genes* all lined up on a chromosome suggests that the *alleles* would assort together, not independently…. but as a **linkage group**
  - Most *alleles* on a chromosome did assort independently, not as linkage groups…. how?
  - Some mechanism to allow neighboring *alleles* to assort independently must exist.



Chromosomes and genome maps

- DNA in most eukarotes is organized into linear chromosomes
  - Having *genes* all lined up on a chromosome suggests that the *alleles* would assort together, not independently…. but as a **linkage group**
  - Most *alleles* on a chromosome did assort independently, not as linkage groups…. how?
  - Some mechanism to allow neighboring *alleles* to assort independently must exist.



Possible chromosomes in gametes

Genetic Mapping

- **Crossing-over** and **recombination**
- Physical distance between loci influences recombination rates between loci
- Recombination fraction is a measure of distance

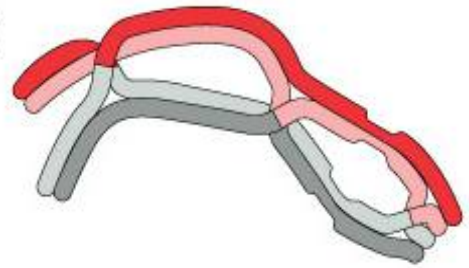
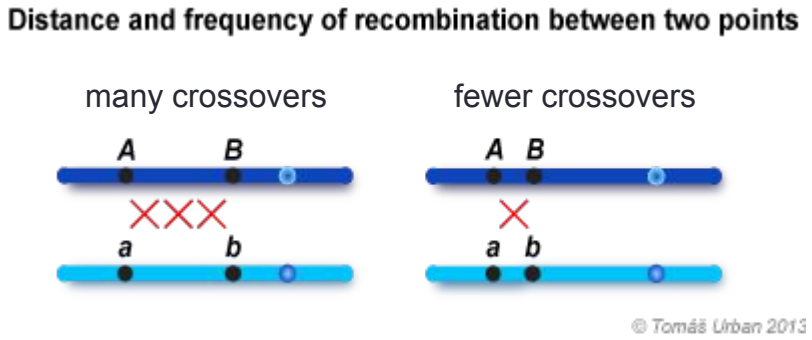
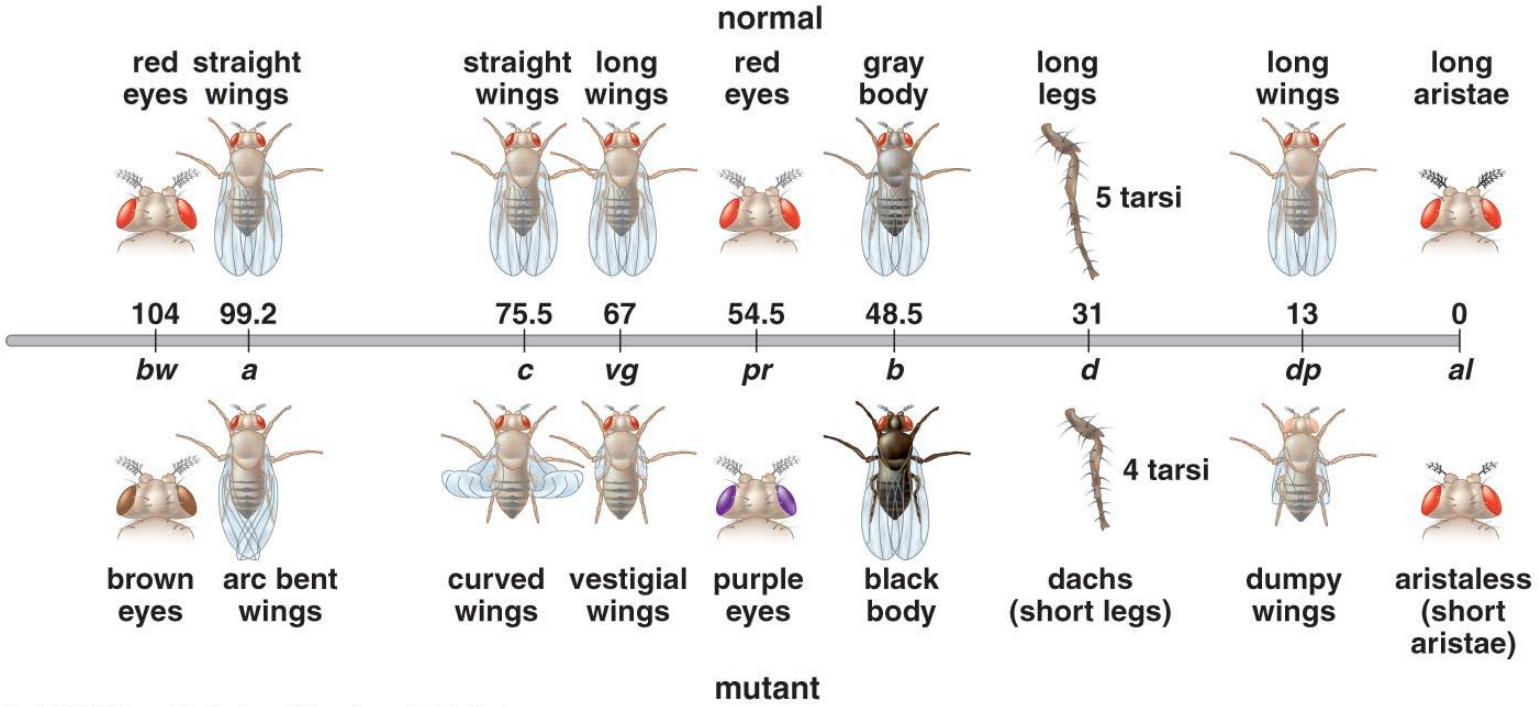


Figure 10-6 Cell and Molecular Biology, 4/e (© 2005 John Wiley & Sons)



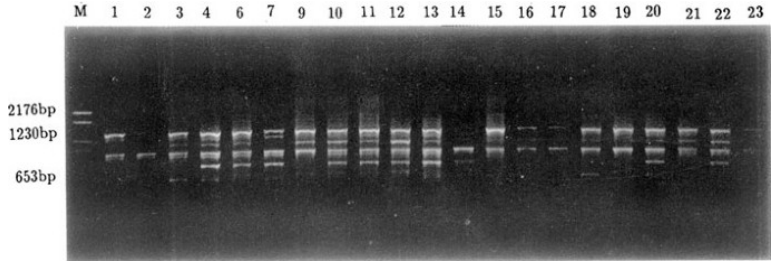
Genetic Mapping

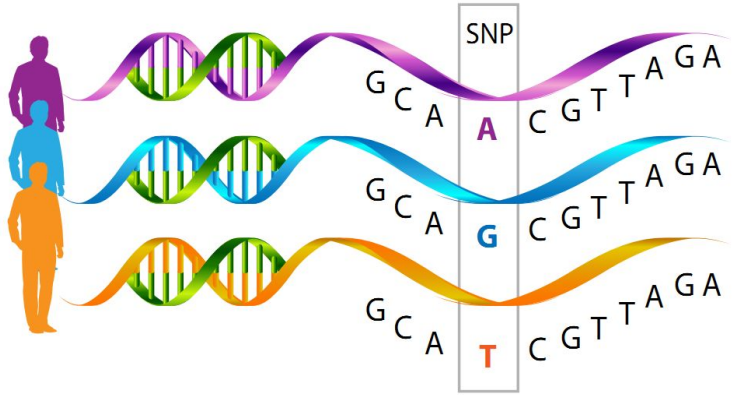
- Thomas Hunt Morgan's lab had recombination fractions for dozens of genes in *Drosophila*
  - Alfred Sturtevant (19-year-old undergrad) developed the world's first genetic map after skipping one nights homework
  - Four linkage groups in Drosophila – four chromosomes

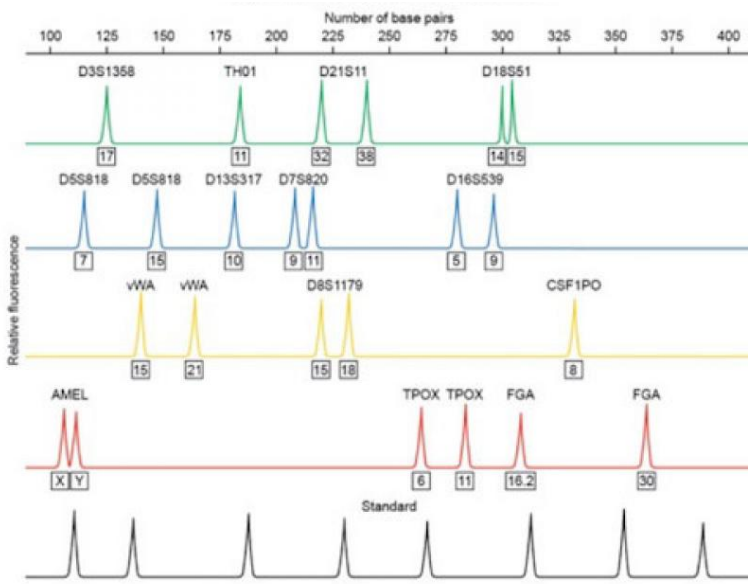


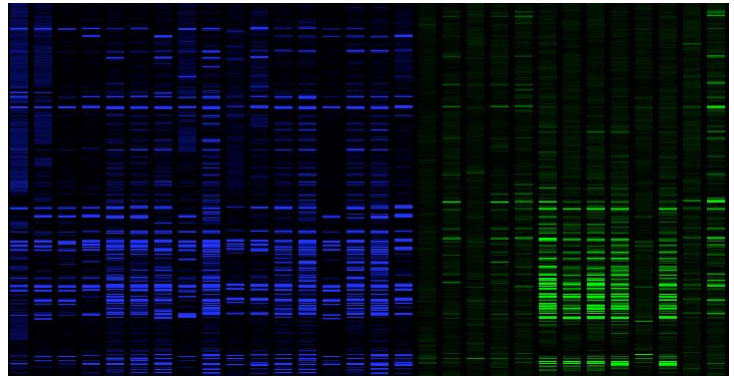
Higher resolution maps

- Phenotype producing genes are few and far between in many eukaryotes (~20k genes total, only a relative few have observable phenotypes)
- More detailed maps rely on DNA sequence variants
  - Restriction digest maps
  - STRs (DNA fingerprint loci)
  - AFLP
  - RAPDs
  - + dozens



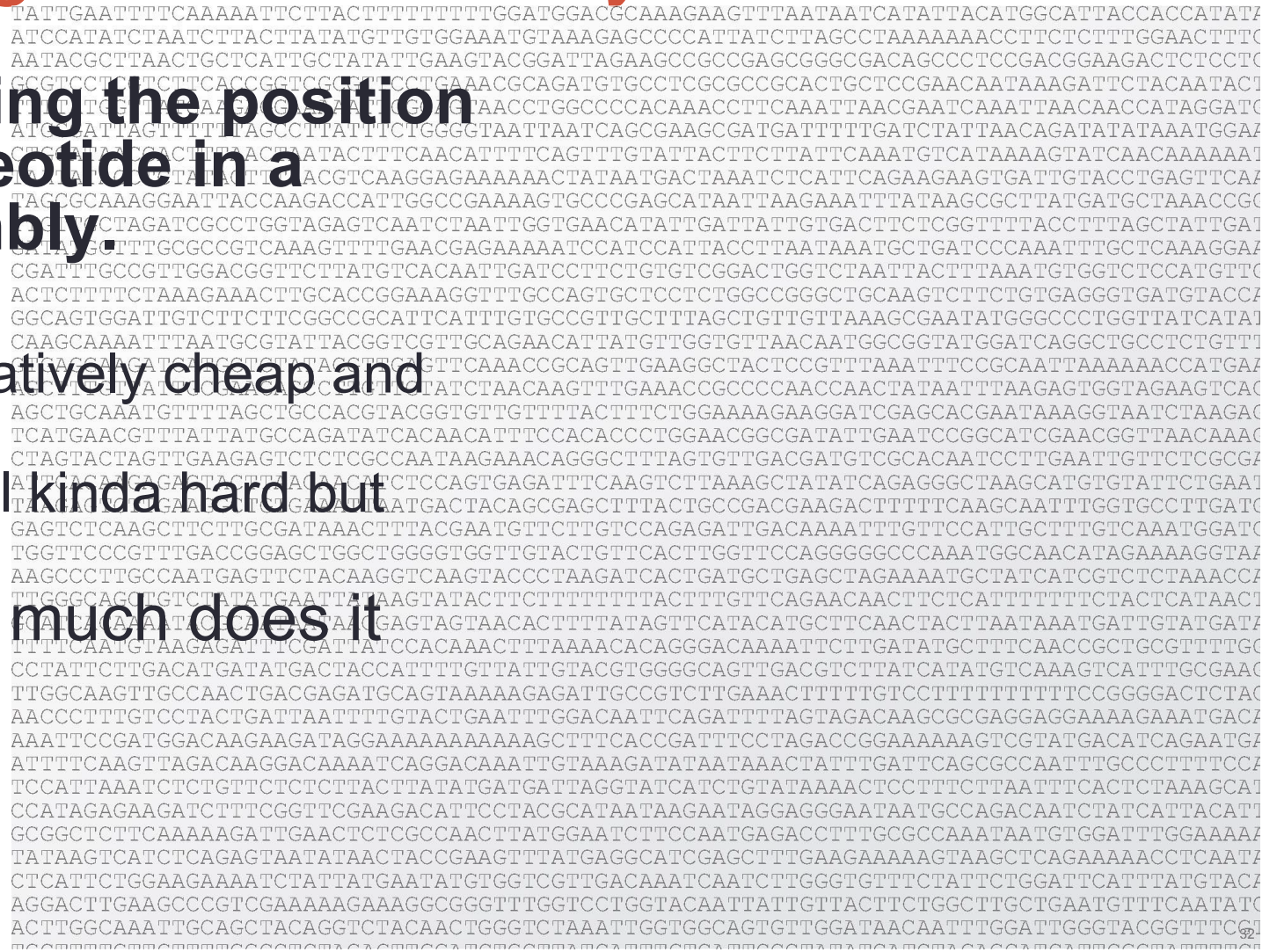






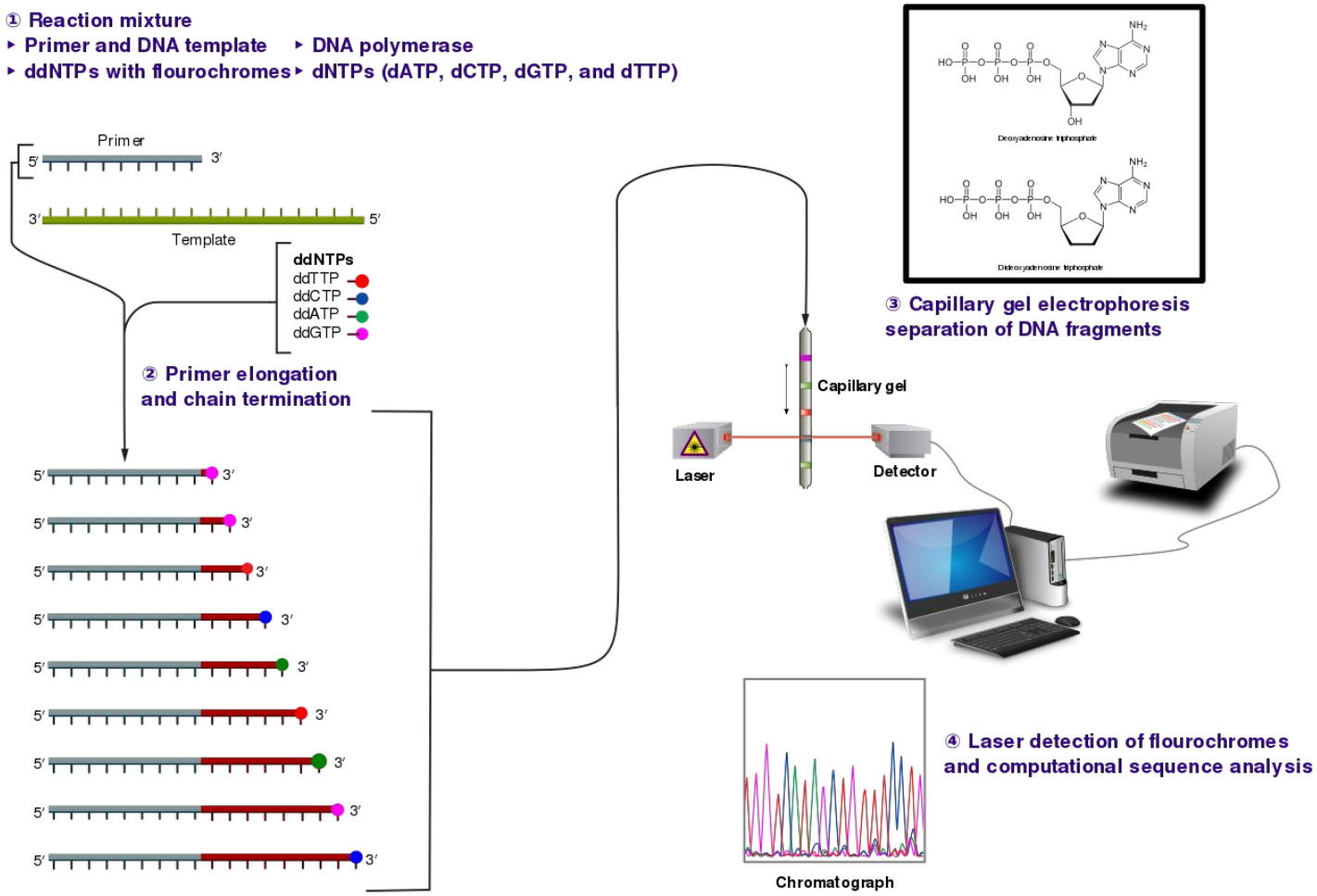
Genome Sequencing and Assembly

- **The ultimate map is knowing the position and identity of every nucleotide in a genome, a genome assembly.**
- Two steps:
  - Sequence the DNA – now relatively cheap and easy
  - Assemble the fragments – still kinda hard but getting easier
- How do we do this and how much does it cost?



DNA sequencing methodologies: ca. 1977
AKA 1st generation sequencing

- Sanger/Chain termination sequencing
  - DNA replication based
  - Substitution of substrate with chain-terminator chemical.
  - Efficient
  - Amenable to automation



DNA Analysis: DNA Sequencing

- Best case scenario for Sanger sequencing
  - 700-1000 bp reads
  - 96 capillary machine
  - 3 hour run time
  - ~700 bp/capillary
  - ~5.4 Mb/day
  - ~600 days to sequence the human genome to 1X depth

DNA Analysis: DNA Sequencing

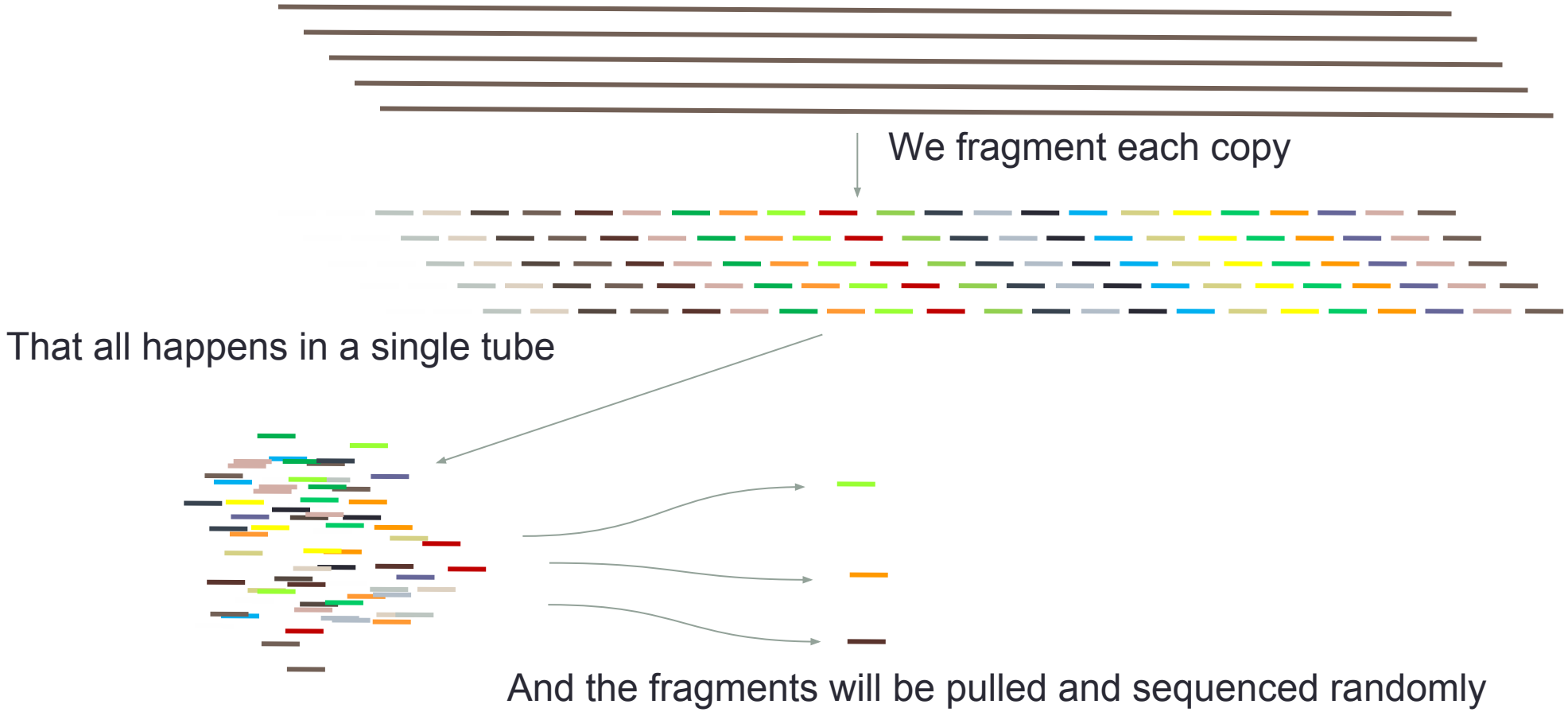
- What is sequencing depth/coverage?
  - Depth is the number of times a given nucleotide has been sequenced, on average
  - ‘Coverage’ is often used interchangeably with ‘depth’
  - Typically, the higher the number, the better the assembly
- 1X coverage = sequencing enough fragments to generate the equivalent of one genome’s worth of sequence data.
  - For the human genome, that would be ~3 billion bases of sequence data
- Is 1X enough to get an accurate genome assembly?

Is 1X enough?

- **No.** Why?
  - No current sequencing technologies allow for telomere-to-telomere sequence reads
  - Sequence reads that are relatively small compared to chromosome length. We stitch those randomly selected reads together to assemble them
  - We also almost always rely on making our sequencing libraries from multiple copies of the species' genome
  - Random selection of fragments to sequence
  - All sequencing technologies involve error

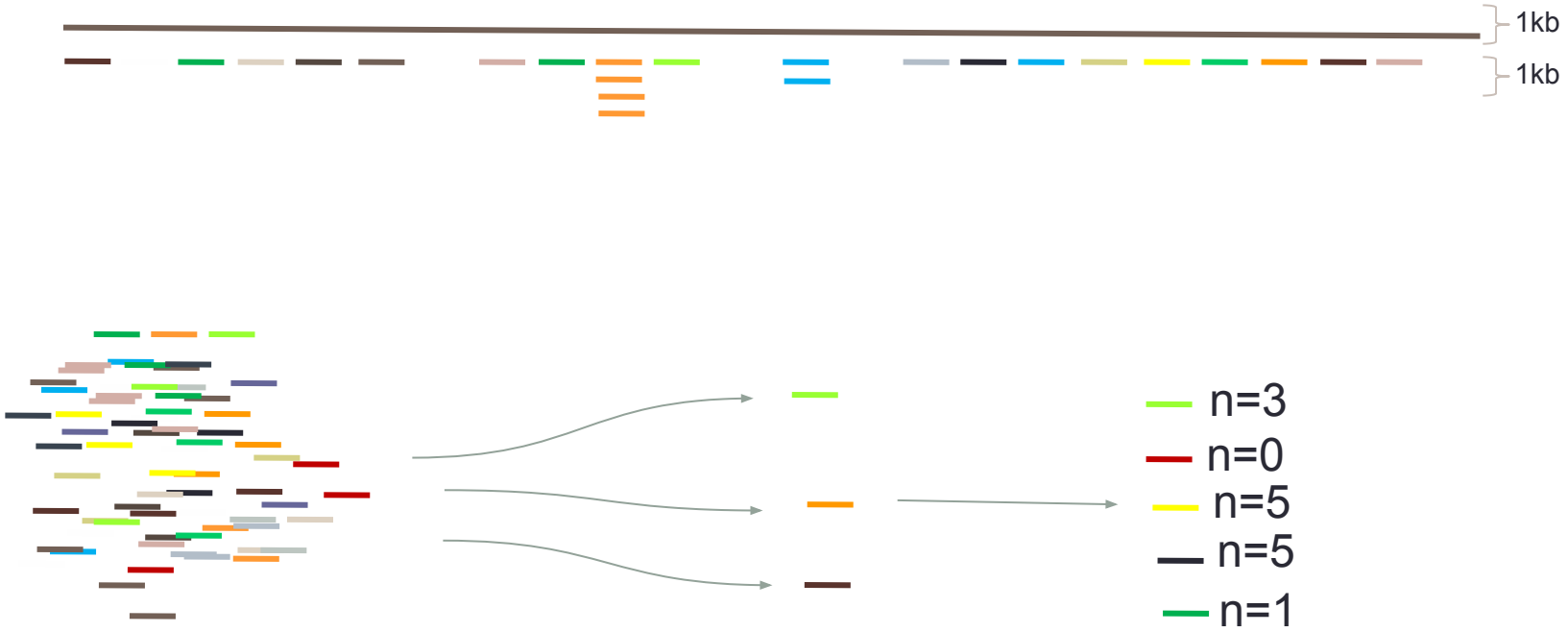
Is 1X enough?

- Imagine you have a genome consisting of only one chromosome and have isolated five cells worth of DNA from that organism.



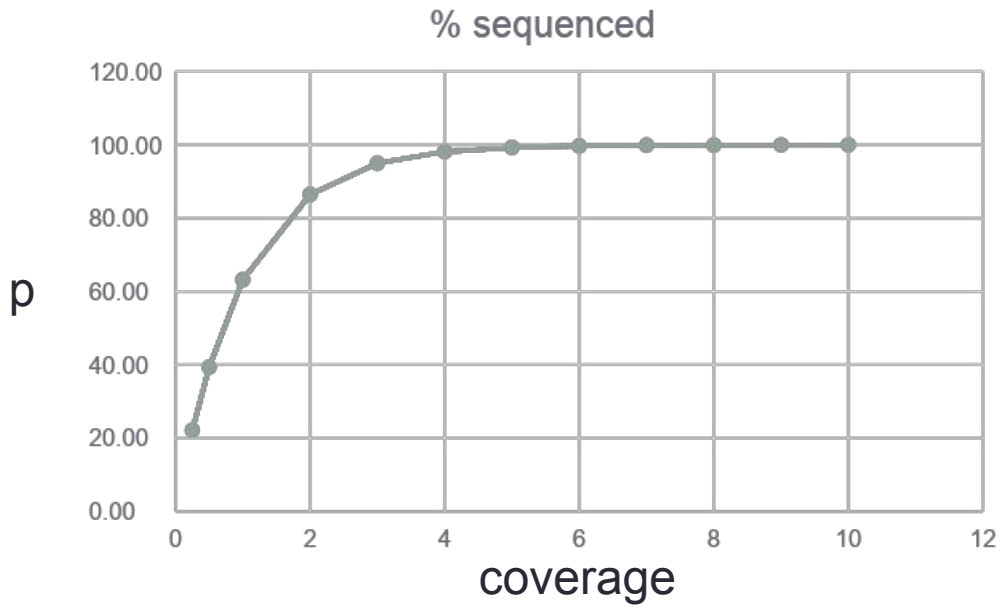
Is 1X enough?

- We'll get some fragments >1 time, some a bunch of times, and some not at all.



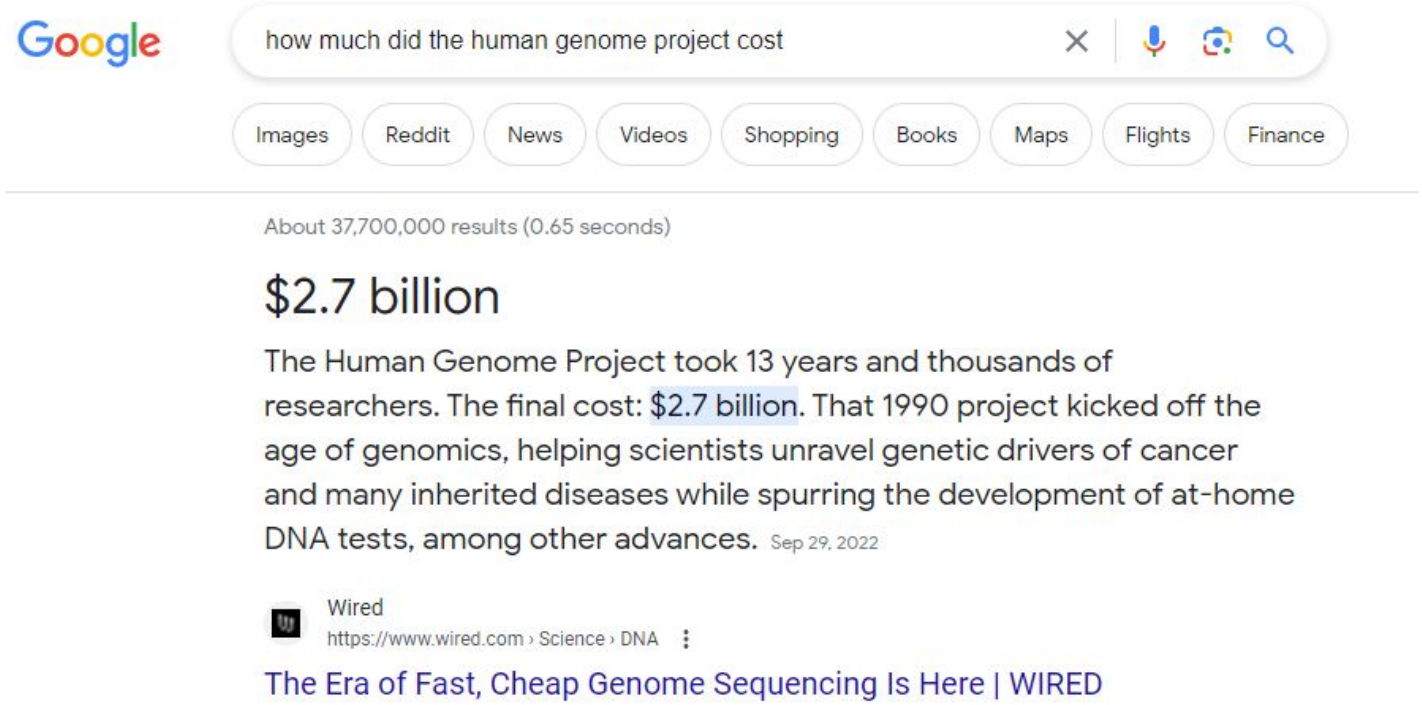
Is 1X enough?

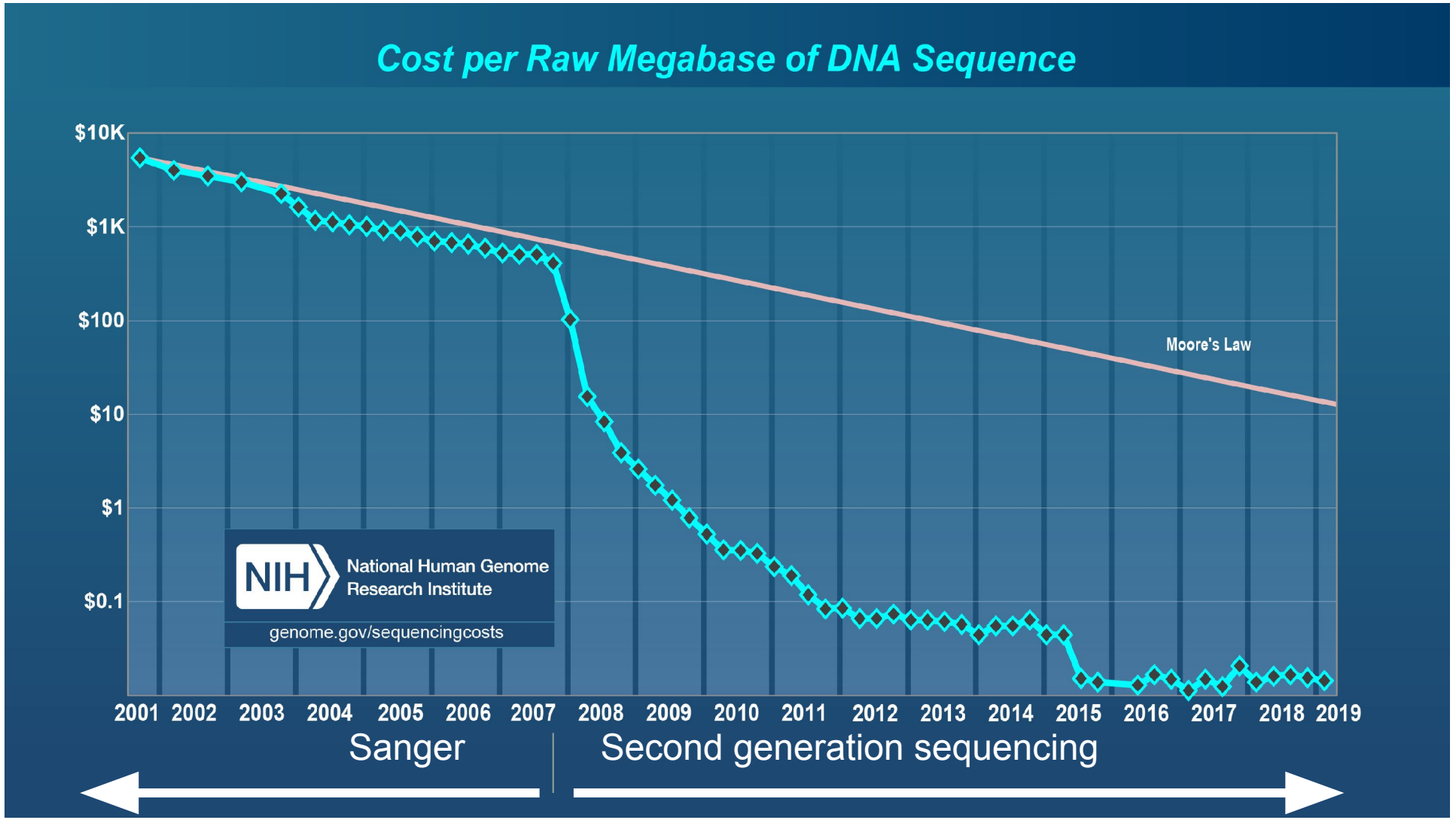
- **Probability of a base being sequenced (p) = $1-e^{-c}$**
- **C = Coverage**
- **At 1x, p = $1-e^{-1}$ = 63.21**
- **<u>We'll miss ~36.79 % of the genome on average</u>**



Genome Assembly

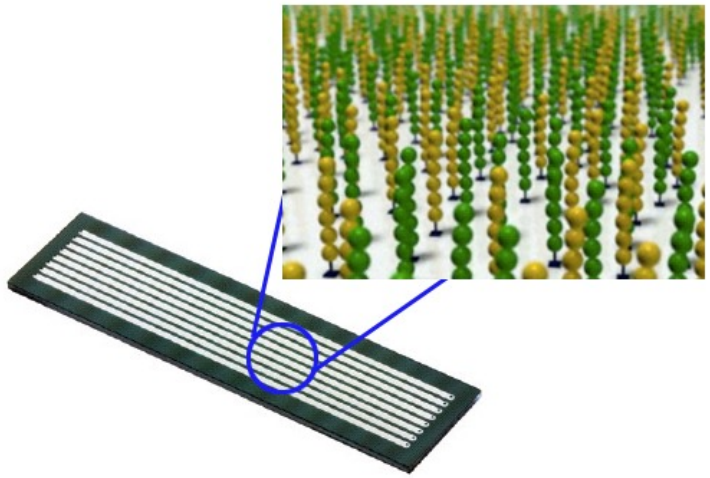
- **Cost/Mb (1990s, Sanger sequencing) = ~$9000**
- **Human genome = 3000Mb**
- **Cost/1X human genome = $27,000,000**
- **Cost/10X human genome = $270,000,000**
- **But that's just the sequencing**

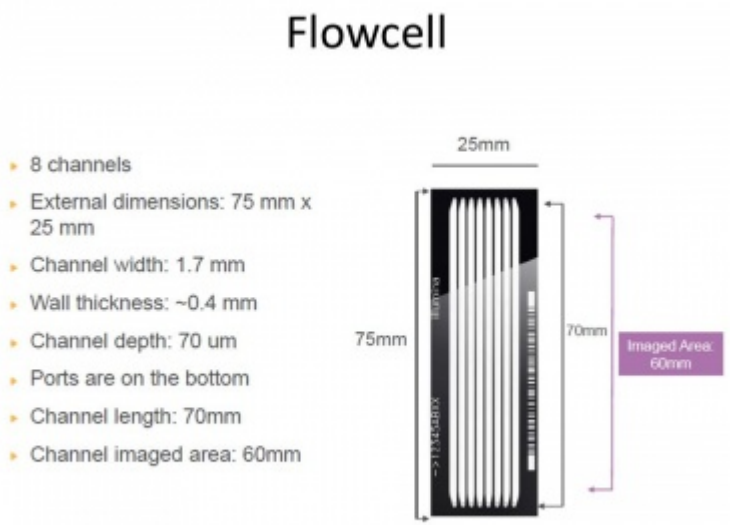


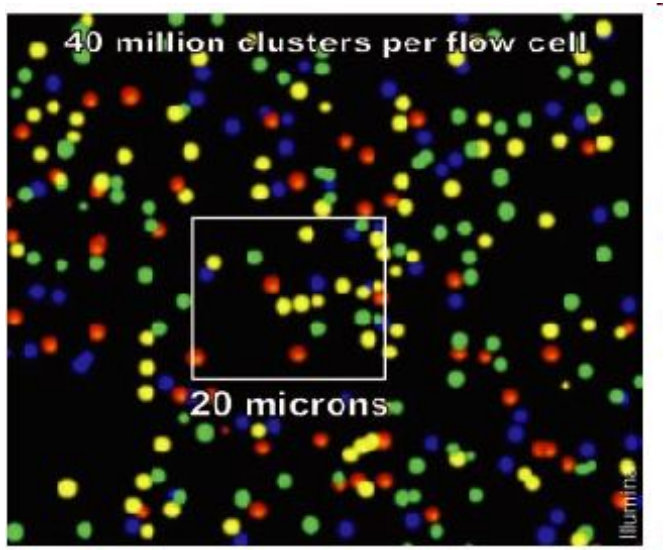

Cost per Raw Megabase of DNA Sequence
$10K
$1K
$100
$10
$1
$0.1
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
Sanger
Second generation sequencing

Next (2nd) Generation Sequencing Illumina

- **Illumina** is a massively-parallel sequencing-by-synthesis system
- Really took off ~2007
  - Currently dominates 2nd gen
- Works using flow cells





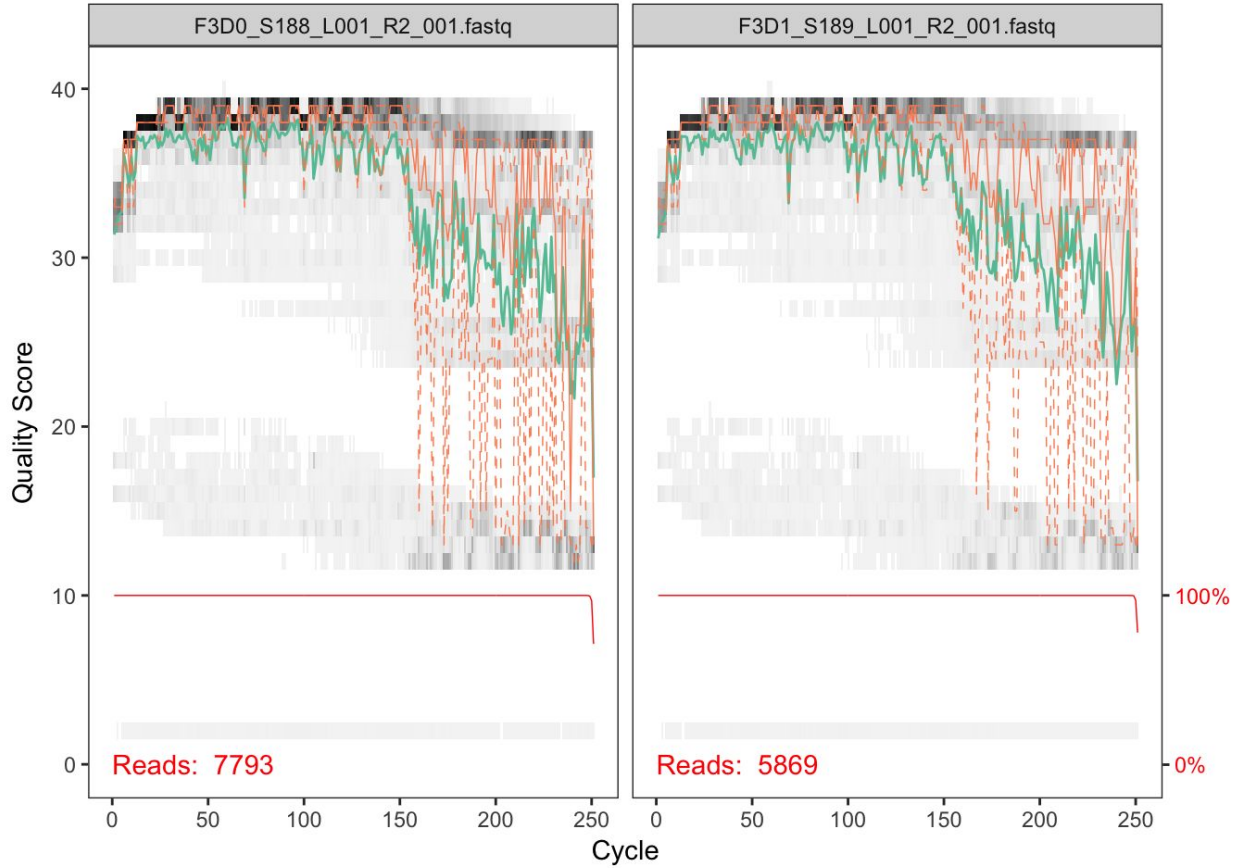


Must see video - https://www.youtube.com/watch?v=fCd6B5HRaZ8

Next (2nd) Generation Sequencing
Illumina

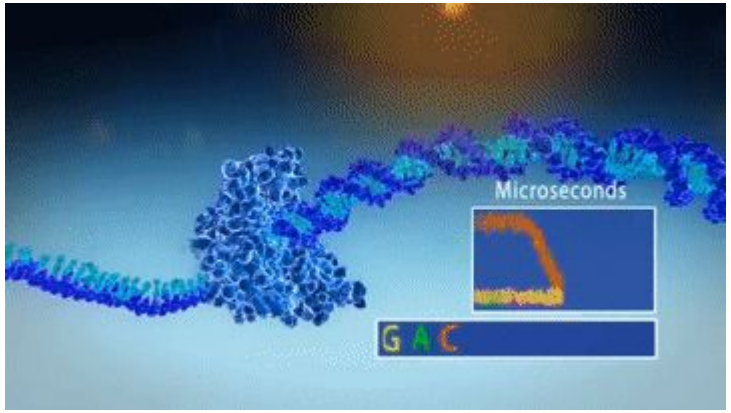
- Pros
  - Illumina sequences are highly accurate (99.9+%)
  - Cheap
  - Library prep is reasonably easy
- Cons
  - Maximum read length is ~300 bp

- **Cost/Gb = $4-$200**
- **Human genome = 3Gb**
- **Cost/1X human genome = $12**
- **Cost/100X human genome = $120**

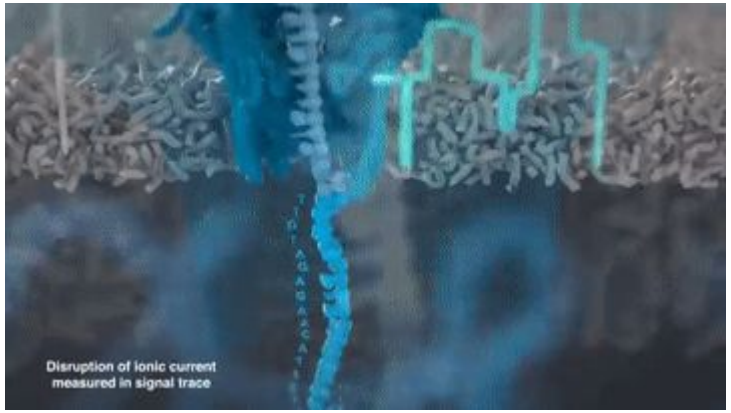


Next (3rd) Generation Sequencing PacBio and Nanopore

- Read lengths 10-40 kb on average
- Performance
  - Long reads – 10 - 40kb on average
  - <1 day run time



https://www.youtube.com/watch?v=v8p4ph2MAvI

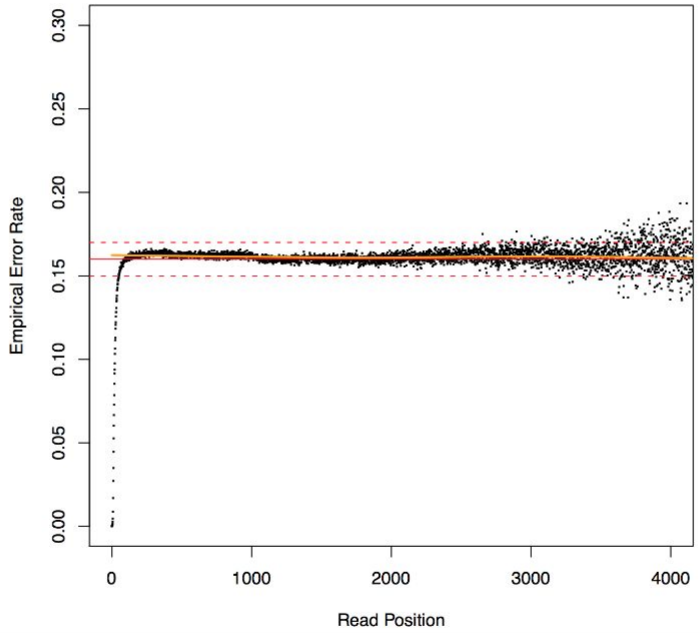


https://www.youtube.com/watch?v=RcP85JHLmnI

Next (3nd) Generation Sequencing
PacBio and Nanopore

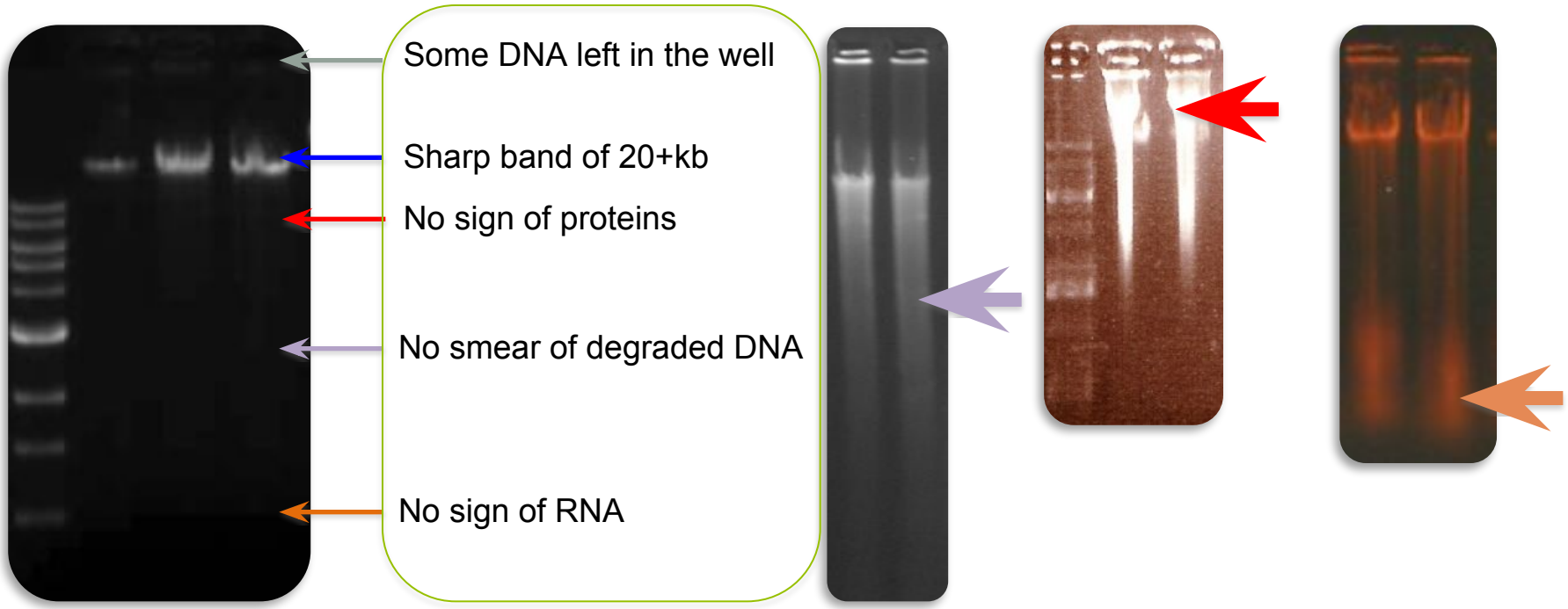
- Pros
  - Cheap but not as cheap
  - Library prep is reasonably easy
  - Long reads
- Cons
  - More expensive
  - Higher error rate

- Error profile
- ~15% as of 2012
- ~13% as of 2017



- **Cost/Gb = $30**
- **Human genome = 3Gb**
- **Cost/1X human genome = $90**
- **Cost/100X human genome = $900**

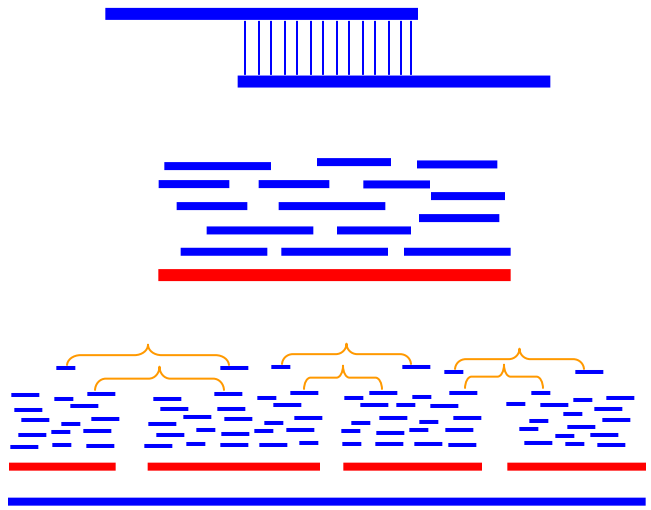
Long read technologies require pure, high molecular weight DNA



- This requires special handling of the tissues prior to DNA extraction
- This can be *logistically difficult* and *cost-prohibitive*.

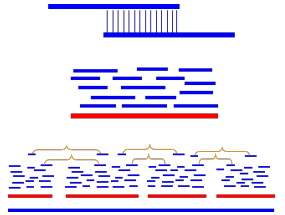
Genome Assembly

- **Reminders**
  - Getting the sequence is just the first step
  - Reads (even long reads) are very short compared to the length of an entire genome
- **De novo genome assembly**
  - Obtain reads
  - Assemble reads into contigs
  - Assemble contigs into scaffolds
  - Refine scaffolds into a consensus sequence
  - If possible, map scaffolds to chromosomes



..ACGATTACAATAGGTT..

An assembly problem



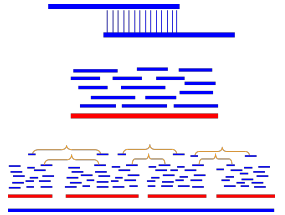
Single-end reads

```
tttggtcgca      gtcgccccta
tccagcatca      cggctagata
ccagcaccac      cacttagccg
ggctagatat      tttggtcgcc
tcacttagcc      cccctatacg
gccgaaaaag
```

Paired-end reads

```
tgggtgt----------cgaaaaa
ggtagcc----------attgggt
tttggtc----------gatattg
agatatt----------cttagct
```

An assembly problem



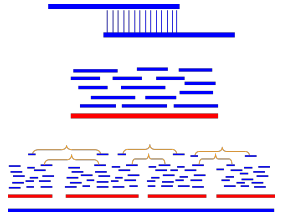
Single-end reads

tttggtcgca
tccagcatca
ccagcaccac
ggctagatat
tcacttagcc
gccgaaaaag
gtcgccccta
cggctagata
cacttagccg
tttggtcgcc
cccctatacg

Paired-end reads

tgggtgt----------cgaaaaa
ggtagcc----------attgggt
tttggtc----------gatattg
agatatt----------cttagct

An assembly problem



```
tttggtcgca
tttggtcgcc
```

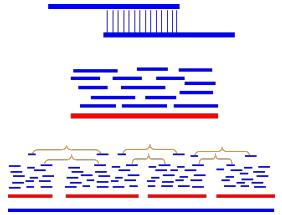
Single-end reads

```
                gtcgccccta
tccagcatca      cggctagata
ccagcaccac      cacttagccg
ggctagatat      ccccctatacg
tcacttagcc
gccgaaaaag
```

Paired-end reads

```
tgggtgt----------cgaaaaa
ggtagcc----------attgggt
tttggtc----------gatattg
agatatt----------cttagct
```

An assembly problem



```
tttggtcgca
tttggtcgcc

gtcgccccta
    cccctatacg
```

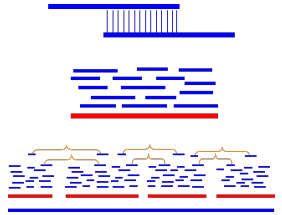
Single-end reads

```
tccagcatca
ccagcaccac
ggctagatat
tcacttagcc
gccgaaaaag

cggctagata
cacttagccg
```

Paired-end reads

```
tgggtgt----------cgaaaaa
ggtagcc----------attgggt
tttggtc----------gatattg
agatatt----------cttagct
```

An assembly problem



```
tttggtcgca
tttggtcgcc

gtcgccccta
    cccctatacg
```

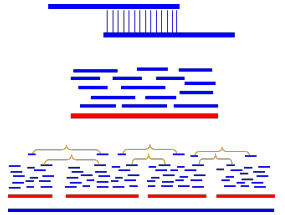
Single-end reads

```
tccagcatca
ccagcaccac
ggctagatat
tcacttagcc
gccgaaaaag

cggctagata
cacttagccg
```

Paired-end reads

```
tgggtgt----------cgaaaaa
ggtagcc----------attgggt
tttggtc----------gatattg
agatatt----------cttagct
```

An assembly problem



```
tttggtcgca
tttggtcgcc

gtcgccccta
    ccccctatacg

tcacttagcc
 cacttagccg


cggctagata
 ggctagatat
```

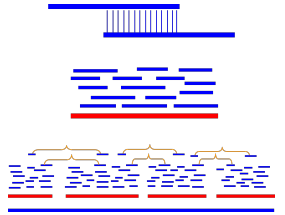
Single-end reads

```
tccagcatca
ccagcaccac


gccgaaaaag
```

Paired-end reads

```
tgggtgt---------------cgaaaaa
ggtagcc---------------attgggt
tttggtc---------------gatattg
agatatt---------------cttagct
```

An assembly problem

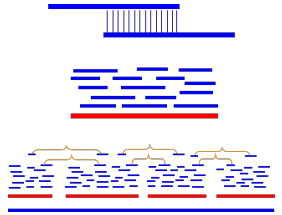


```
tttggtcgca
tttggtcgcc

gtcgccccta
    cccctatacg

tcacttagcc
 cacttagccg


cggctagata
 ggctagatat

tccagcatca
 ccagcaccac


gccgaaaaag
```

Paired-end reads

```
tgggtgt---------------cgaaaaa
ggtagcc---------------attgggt
tttggtc---------------gatattg
agatatt---------------cttagct
```

An assembly problem

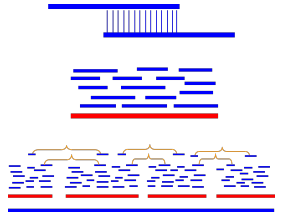


```
tttggtcgca
tttggtcgcc

gtcgccccta
    ccccctatacg

tcacttagcc
 cacttagccg


cggctagata
 ggctagatat

tccagcatca
 ccagcaccac


gccgaaaaag
```

Paired-end reads

```
tgggtgt---------------cgaaaaa
ggtagcc---------------attgggt
tttggtc---------------gatattg
agatatt---------------cttagct
```

An assembly problem



```
tttggtcgca
tttggtcgcc

gtcgccccta
    cccctatacg
```

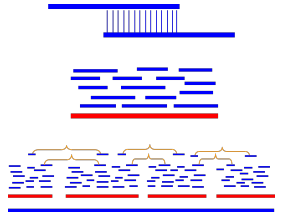
```
cggctagata
 ggctagatat

tccagcatca
 ccagcaccac
```

```
tcacttagcc
 cacttagccg
      gccgaaaaag
```

Paired-end reads

```
tgggtgt---------------cgaaaaa
ggtagcc---------------attgggt
tttggtc---------------gatattg
agatatt---------------cttagct
```

An assembly problem



```
tttggtcgca
tttggtcgcc

gtcgccccta
    cccctatacg
```

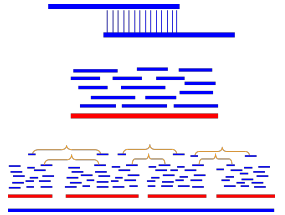
```
cggctagata
 ggctagatat

tccagcatca
 ccagcaccac
```

```
tcacttagcc
 cacttagccg
       gccgaaaaag
```

Paired-end reads

```
tgggtgt---------------cgaaaaa
ggtagcc---------------attgggt
tttggtc---------------gatattg
agatatt---------------cttagct
```

An assembly problem



```
tttggtcgca
tttggtcgcc

gtcgccccta
    ccccctatacg
```

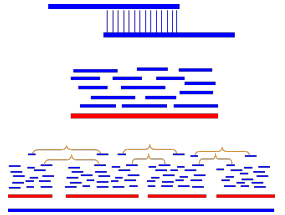
```
cggctagata
 ggctagatat

tccagcatca
 ccagcaccac
```

```
tcacttagcc
 cacttagccg
       gccgaaaaag
```

Paired-end reads

```
tgggtgt---------------cgaaaaa
ggtagcc---------------attgggt
tttggtc---------------gatattg
agatatt---------------cttagct
```

An assembly problem

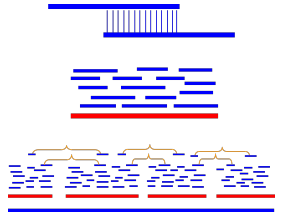


```
                    tttggtcgca
                    tttggtcgcc

gtcgccccta
     cccctatacg

cggctagata
 ggctagatat

tccagcatca
 ccagcaccac

                          tcacttagcc
                           cacttagccg
                                 gccgaaaaag
```

Paired-end reads

```
tgggtgt---------------cgaaaaa
ggtagcc---------------attgggt
tttggtc---------------gatattg
agatatt---------------cttagct
```

An assembly problem



tttggtcgca

tttggtcgcc

gtcgccccta

ccctatacg

tcacttagcc

cacttagccg

gccgaaaaag

cggctagata

ggctagatat

tccagcatca

ccagcaccac

Paired-end reads

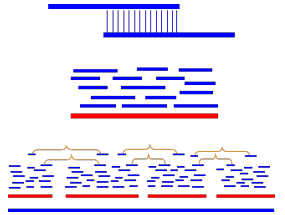
tgggtgt----------cgaaaaa

ggtagcc----------attgggt

tttggtc----------gatattg

agatatt----------cttagct

An assembly problem



```
tttggtcgca
tttggtcgcc
    gtcgccccta
        cccctatacg
```

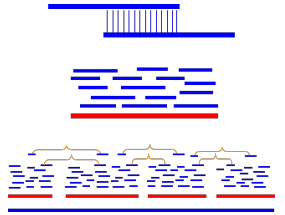
```
tcacttagcc
 cacttagccg
       gccgaaaaag
```

```
cggctagata
 ggctagatat

tccagcatca
 ccagcaccac
```

Paired-end reads

```
tgggtgt---------------cgaaaaa
ggtagcc---------------attgggt
tttggtc---------------gatattg
agatatt---------------cttagct
```

An assembly problem



```
tttggtcgca
tttggtcgcc
    gtcgccccta
        cccctatacg
```

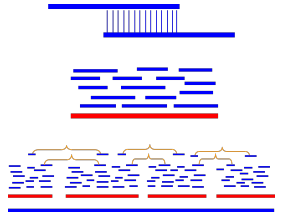
```
tcacttagcc
 cacttagccg
       gccgaaaaag
```

```
cggctagata
 ggctagatat

tccagcatca
 ccagcaccac
```

Paired-end reads

```
tgggtgt---------------cgaaaaa
ggtagcc---------------attgggt
tttggtc---------------gatattg
agatatt---------------cttagct
```

An assembly problem



```
cggctagata
 ggctagatat
```

```
tttggtcgca
tttggtcgcc
    gtcgccccta
        cccctatacg
```

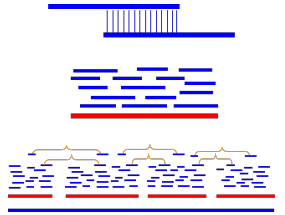
```
tcacttagcc
 cacttagccg
       gccgaaaaag
```

```
tccagcatca
 ccagcaccac
```

Paired-end reads

```
tgggtgt---------------cgaaaaa
ggtagcc---------------attgggt
tttggtc---------------gatattg
agatatt---------------cttagct
```

An assembly problem

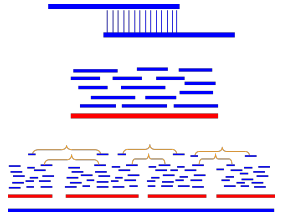


```
                                  cggctagata
                                   ggctagatat
                                  cggctagatat

tttggtcgca
tttggtcgcc
    gtcgccccta
        cccctatacg

tttggtcgcccctatacg

                                          tcacttagcc
                                           cacttagccg
                                  tccagcatca     gccgaaaaag
                                   ccagcaccac
                                  tccagcatcacttagccgaaaaag
```

Paired-end reads

```
tgggtgt----------------cgaaaaa
ggtagcc----------------attgggt
tttggtc----------------gatattg
agatatt----------------cttagct
```

An assembly problem



tttggtcgca

tttggtcgcc

gtcgccccta

cccctatacg

cggctagata

ggctagatat

tcacttagcc

cacttagccg

tccagcatca gccgaaaaag

ccagcaccac

cggctagatat

tttggtcgcccctatacg

tccagcatcacttagccgaaaaag

Paired-end reads

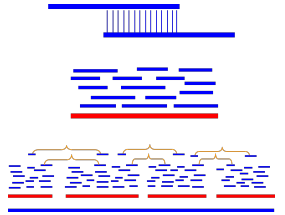
tgggtgt----------cgaaaaa

ggtagcc----------attgggt

tttggtc----------gatattg

agatatt----------cttagct

An assembly problem



tttggtcgca

tttggtcgcc

gtcgccccta

cccctatacg

cggctagata

ggctagatat

tcacttagcc

cacttagccg

tccagcatca

gccgaaaaag

ccagcaccac

cggctagatat

tttggtcgcccctatacg

tccagcatcacttagccgaaaaag

Paired-end reads

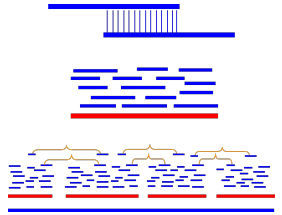
tgggtgt----------cgaaaaa

ggtagcc----------attgggt

tttggtc----------gatattg

agatatt----------cttagct

An assembly problem



tttggtcgca

tttggtcgcc

gtcgccccta

cccctatacg

cggctagata

ggctagatat

tcacttagcc

cacttagccg

tccagcatca gccgaaaaag

ccagcaccac

cggctagatat

tttggtcgcccctatacg

tttggtc----------gatattg

tccagcatcacttagccgaaaaag

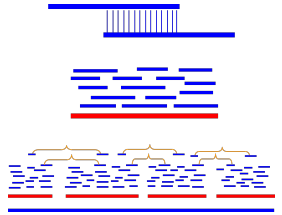
Paired-end reads

tgggtgt----------cgaaaaa

ggtagcc----------attgggt

agatatt----------cttagct

An assembly problem



tttggtcgca

tttggtcgcc

gtcgccccta

cccctatacg

cggctagata

ggctagatat

tcacttagcc

cacttagccg

tccagcatca gccgaaaaag

ccagcaccac

cggctagatat

tttggtcgcccctatacg

tttggtc---------------gatattg

tccagcatcacttagccgaaaaag

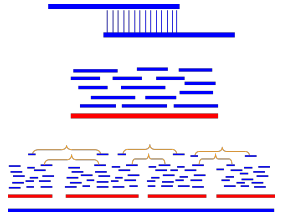
Paired-end reads

tgggtgt---------------cgaaaaa

ggtagcc---------------attgggt

agatatt---------------cttagct

An assembly problem



```
tttggtcgca
tttggtcgcc
    gtcgccccta
        cccctatacg
```

```
cggctagata
 ggctagatat
```

```
tcacttagcc
 cacttagccg
tccagcatca    gccgaaaaag
 ccagcaccac
```

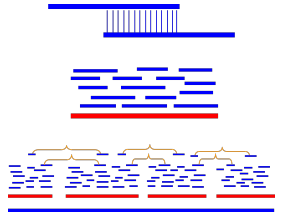
```
                cggctagatat
tttggtcgcccctatacg
tttggtc---------------gatattg
```

```
tccagcatcacttagccgaaaaag
```

Paired-end reads

```
tgggtgt---------------cgaaaaa
ggtagcc---------------attgggt

agatatt---------------cttagct
```

An assembly problem



tttggtcgca

tttggtcgcc

gtcgccccta

cccctatacg

cggctagata

ggctagatat

tcacttagcc

cacttagccg

tccagcatca gccgaaaaag

ccagcaccac

cggctagatat

tttggtcgcccctatacg

tttggtc----------gatattg

tccagcatcacttagccgaaaaag

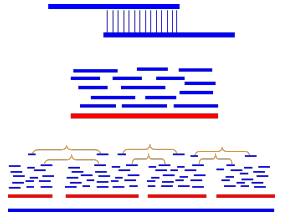
Paired-end reads

tgggtgt----------cgaaaaa

ggtagcc----------attgggt

agatatt----------cttagct

An assembly problem



tttggtcgca
tttggtcgcc
gtcgccccta
cccctatacg

cggctagata
ggctagatat

tcacttagcc
cacttagccg
tccagcatca gccgaaaaag
ccagcaccac

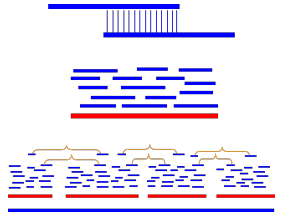
cggctagatat
tttggtcgcccctatacg
tttggtc----------gatattg
agatatt----------cttagct

tccagcatcacttagccgaaaaag

Paired-end reads

tgggtgt----------cgaaaaa
ggtagcc----------attgggt

An assembly problem



tttggtcgca
tttggtcgcc
gtcgccccta
cccctatacg

cggctagata
ggctagatat

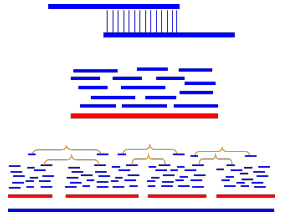
tcacttagcc
cacttagccg
tccagcatca gccgaaaaag
ccagcaccac

cggctagatat
tttggtcgcccctatacg
tttggtc----------gatattg tccagcatcacttagccgaaaaag
agatatt----------cttagct

Paired-end reads

tgggtgt----------cgaaaaa
ggtagcc----------attgggt

An assembly problem



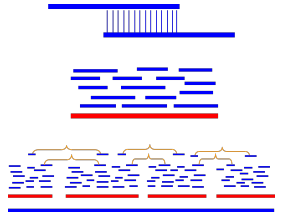
```
tttggtcgca
tttggtcgcc
    gtcgccccta
        cccctatacg
            cggctagata
             ggctagatat
                    tcacttagcc
                     cacttagccg
                tccagcatca    gccgaaaaag
                  ccagcaccac
```

```
                    cggctagatat
tttggtcgcccctatacg
tttggtc---------------gatattg    tccagcatcacttagccgaaaaag
                         agatatt---------------cttagct
```

Paired-end reads

```
tgggtgt---------------cgaaaaa
ggtagcc---------------attgggt
```

An assembly problem



tttggtcgca

tttggtcgcc

gtcgccccta

cccctatacg

cggctagata

ggctagatat

tcacttagcc

cacttagccg

tccagcatca gccgaaaaag

ccagcaccac

cggctagatat

tttggtcgcccctatacg

tttggtc----------gatattg tccagcatcacttagccgaaaaag

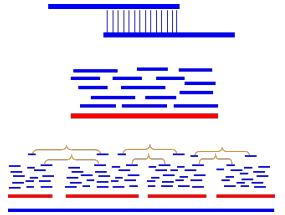
agatatt----------cttagct

tgggtgt----------cgaaaaa

Paired-end reads

ggtagcc----------attgggt

An assembly problem



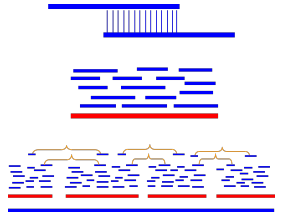
tttggtcgca
tttggtcgcc
gtcgccccta
cccctatacg
cggctagata
ggctagatat
tcacttagcc
cacttagccg
tccagcatca
gccgaaaaag
ccagcaccac

cggctagatat
tttggtcgcccctatacg
tttggtc----------gatattg
tccagcatcacttagccgaaaaag
agatatt----------cttagct
tgggtgt----------cgaaaaa

Paired-end reads

ggtagcc----------attgggt

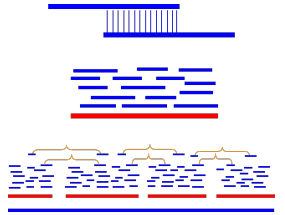
An assembly problem



tttggtcgca
tttggtcgcc
gtcgccccta
cccctatacg
cggctagata
ggctagatat
tcacttagcc
cacttagccg
tccagcatca
gccgaaaaag
ccagcaccac

cggctagatat
tttggtcgcccctatacg
tttggtc----------gatattg
tccagcatcacttagccgaaaaag
agatatt----------cttagct
tgggtgt----------cgaaaaa
ggtagcc----------attgggt

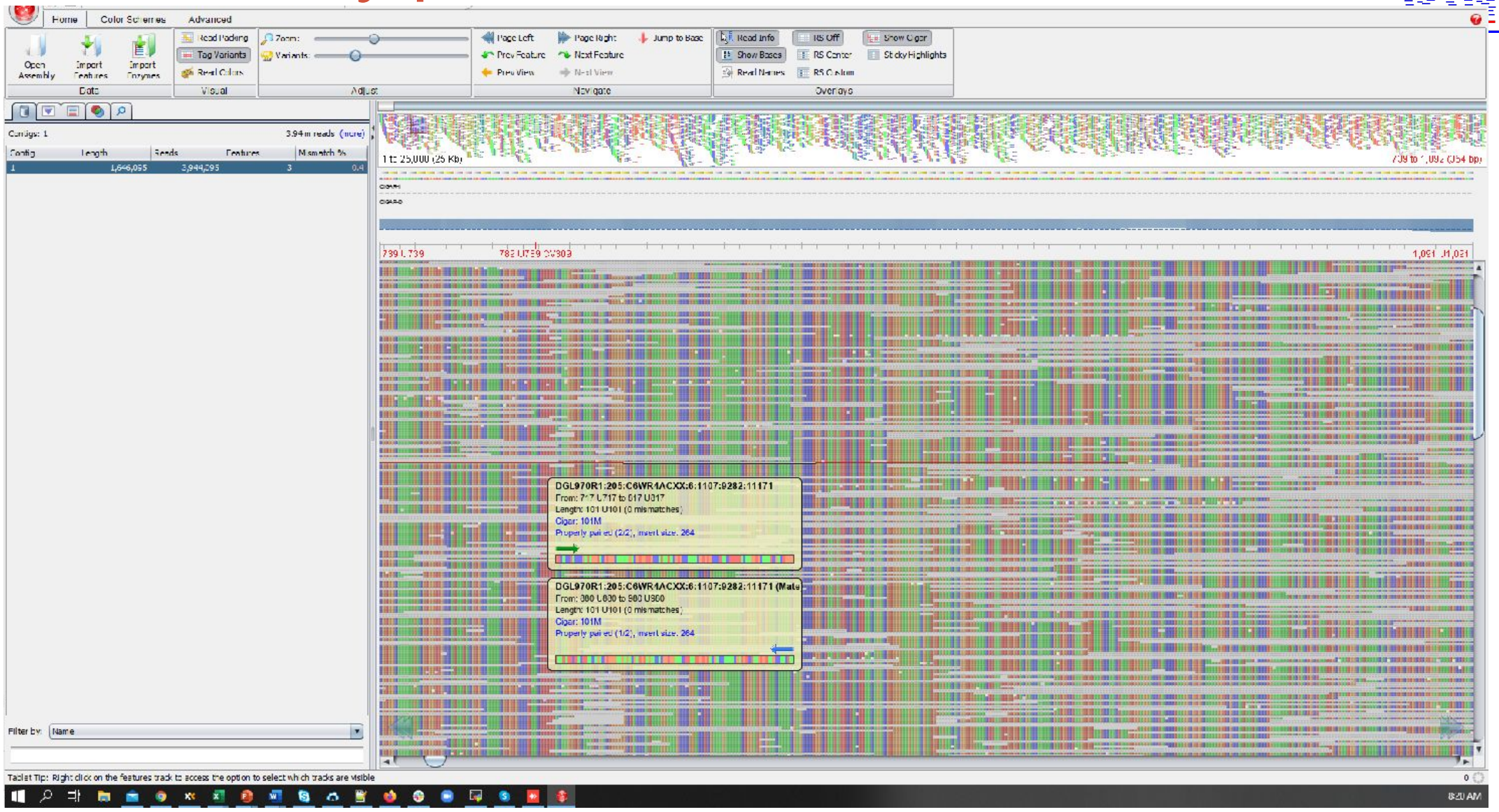
An assembly problem



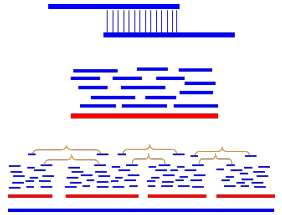
tttggtcgca
tttggtcgcc
gtcgccccta
cccctatacg
cggctagata
ggctagatat
tcacttagcc
cacttagccg
tccagcatca
gccgaaaaag
ccagcaccac

```
                cggctagatat
tttggtcgcccctatacg
tttggtc---------------gatattg    tccagcatcacttagccgaaaaag
                   agatatt---------------cttagct
                        tgggtgt---------------cgaaaaa
   ggtagcc---------------attgggt

tttggtcgcccctatacggctagatattgggtgtccagcatcacttagccgaaaaag
```

An assembly problem



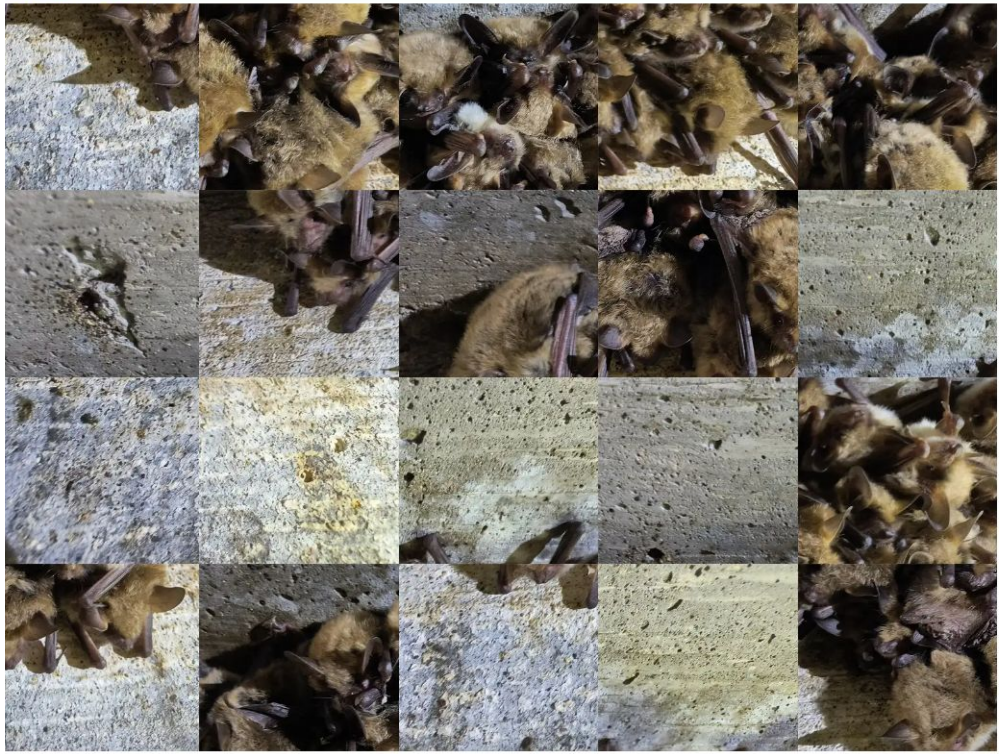
An assembly problem

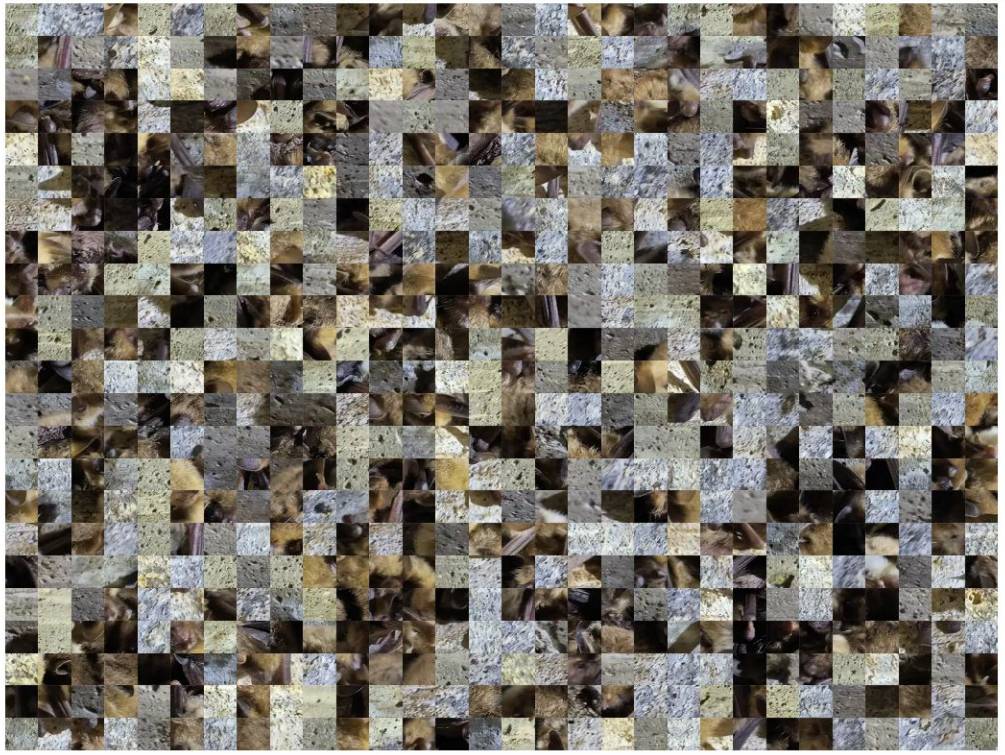


- Sequencing errors

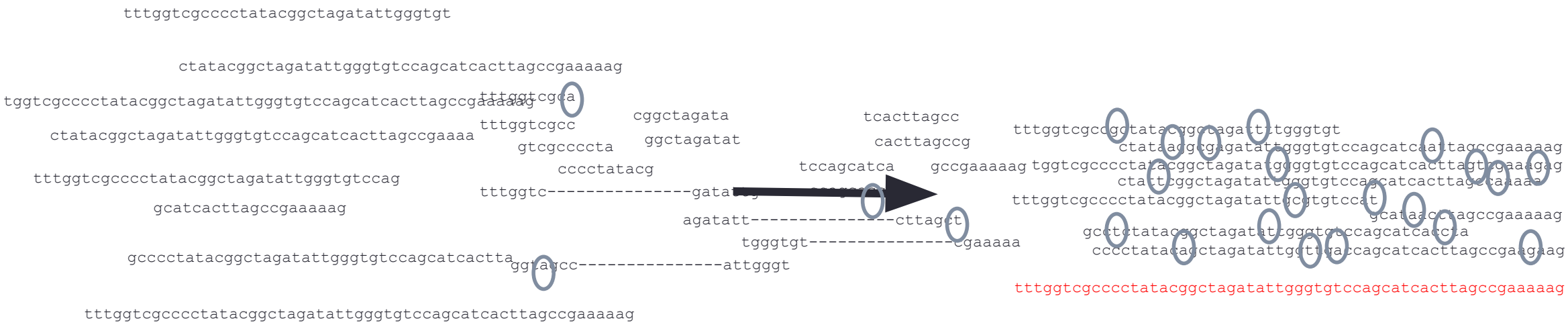
How do long reads help with assembly?

- Which puzzle is easier to solve?



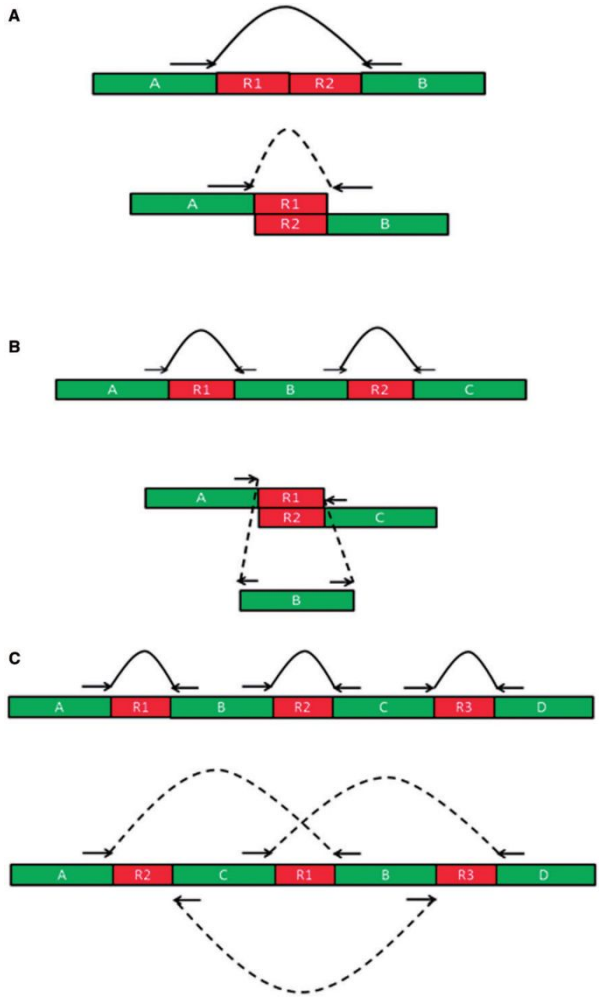


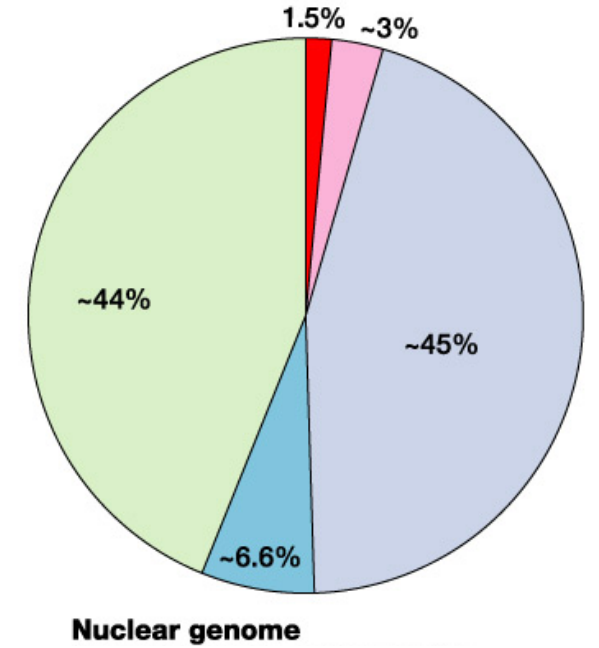
The same assembly problem (but with long reads)



What makes genome assembly difficult?

- Massive scale – billions and billions of reads
- Uneven coverage
- Errors in sequencing reads
- Repetitive regions





Evaluating an assembly

- How do we know if an assembly is any good?
  - Contiguity – How connected are all the pieces?
    - N50 – Minimum scaffold/contig length at which you've covered 50% of the total assembly length
    - Higher value = better
  - Depth
    - Good numbers depend on the sequencing technology
    - Long reads – 50-60X+, Illumina only – 100X+
  - Fidelity – How well does the raw data match the final product?
    - Map the raw reads back to the assembly
    - Higher value = better
  - Gene content – Are the genes we expect to see present in the assembly?
    - BUSCO – Benchmarking Universal Single-Copy Orthologs
    - Higher value = better

Post-assembly applications

- A genome assembly is great but using it to answer biological questions is the real goal
- “What variations in genomes are important for biology?”
- “How does the genome influence biological diversity?”
- Variation can be:
  - **SNPs – single nucleotide polymorphisms**
  - Insertions
  - Deletions
  - Duplications
  - Inversions – >2 bp sequence reversals at particular loci

Detecting Variation

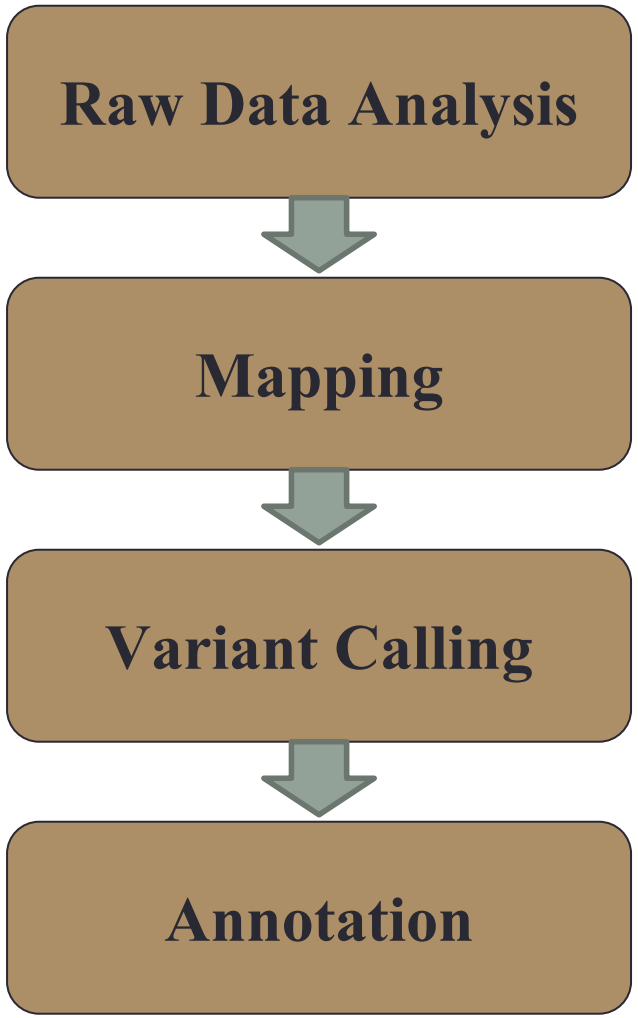


Image processing and base calling

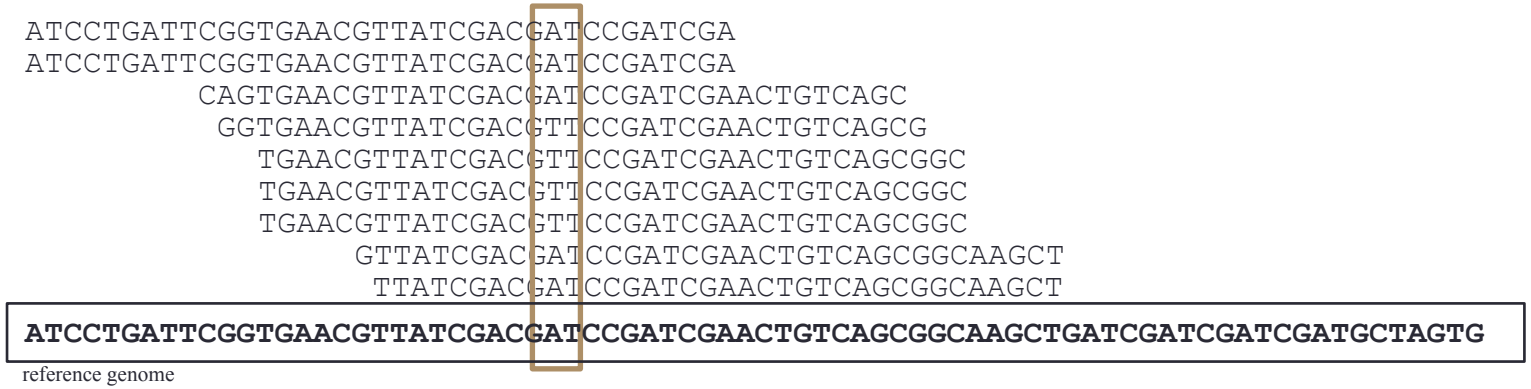
Alignment (to reference genome if available, to each other if not)

Detection of genetic variation (SNPs, Indels, insertions, etc.)

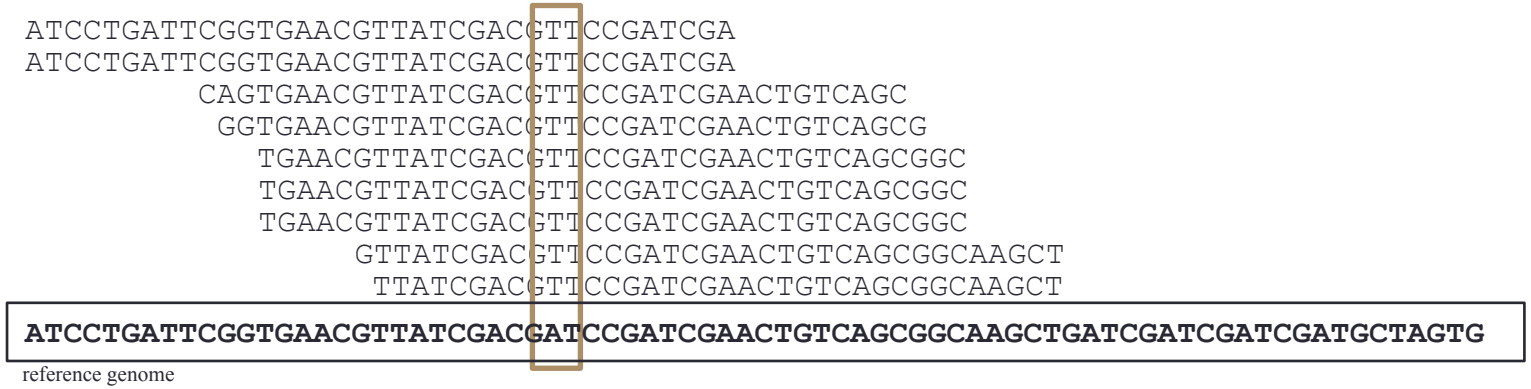
Linking variants to biological information

Read mapping

SNP – heterozygote variant



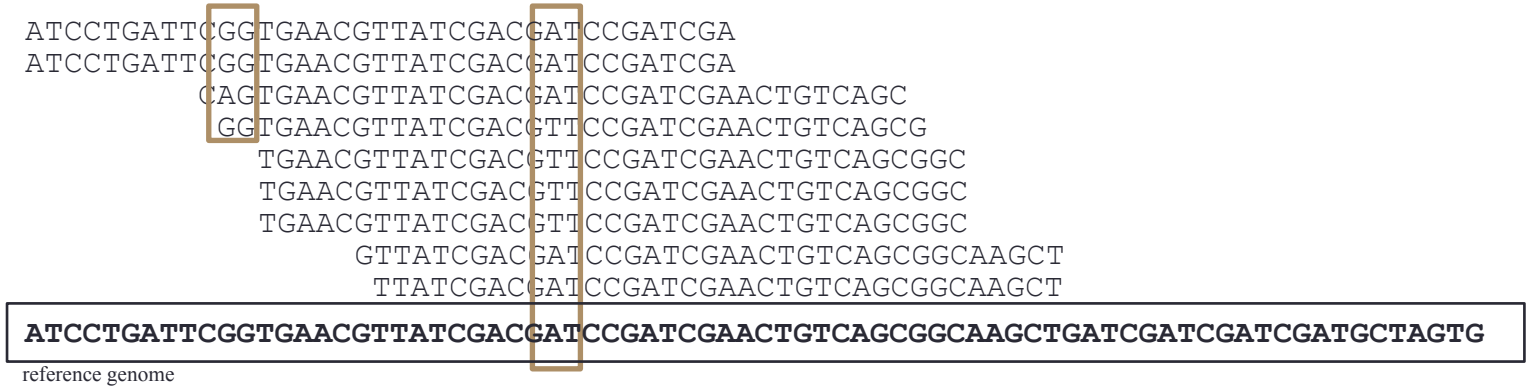
SNP – homozygote variant



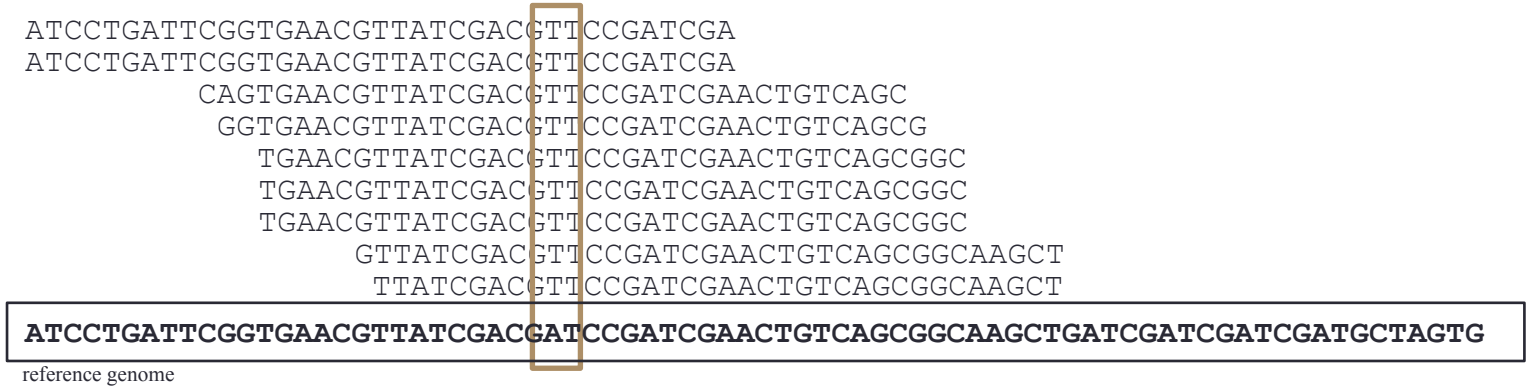
Read mapping

- Appropriate read depth is important to identifying variants

SNP – heterozygote variant



SNP – homozygote variant

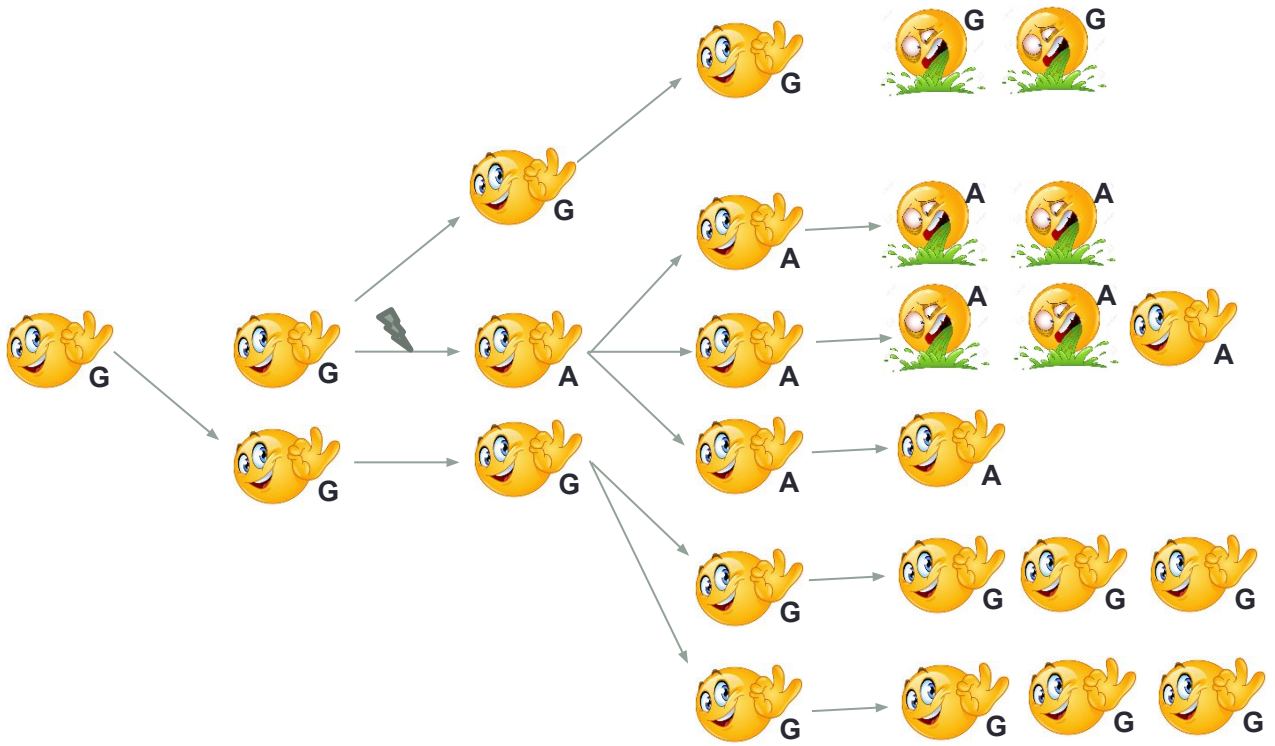


What variants and genes are important to phenotype? - GWAS

- A **genome-wide association study (GWAS) -** identifies SNPs across complete genomes of many samples to find genetic variations statistically associated with a particular phenotype.
- What do you need?
  - An annotated reference genome (usually)
  - A map of genetic variation
  - A set of technologies that can quickly and accurately analyze whole genomes for genetic variants
    - This is typically accomplished using low coverage sequencing (4-20X).

The basic idea

The A allele is associated (4/14, 29%)
with individuals exhibiting the disease phenotype



GWAS and age-related macular degeneration

- Study cohort – 2172 unrelated individuals of European descent, at least 60 years old
  - 1238 with AMD, 934 controls
  - Each individual harbors two alleles
  - 2476 AMD alleles
  - 1868 non-AMD alleles
- Null hypothesis – Alleles will be randomly distributed in the population, i.e. no association of any alleles with AMD
- Alternative hypothesis – Some allele will be positively associated with AMD

934 controls

1238 cases

Age-related macular degeneration

- Single SNP identified by GWAS, rs1061170
- 4344 alleles recovered, two variants C/T

Allele	Cases with AMD	Controls	Total Alleles
C	1522	670	2192
T	954	1198	2152
Total alleles	2476	1868	4344

- X^2 test suggests association, $p=1.2 \times 10^{-62}$

Evolutionary history - *Myotis* phylogenomics

SINE-Based Phylogenomics Reveal Extensive Introgression and Incomplete Lineage Sorting in *Myotis*

by Jennifer M. Korstian [1], Nicole S. Paulat [1], Roy N. Platt II [2], Richard D. Stevens [3] and David A. Ray [1,*]

[1] Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA

[2] Host-Pathogen Interactions Program, Texas Biomedical Research Institute, San Antonio, TX 78227, USA

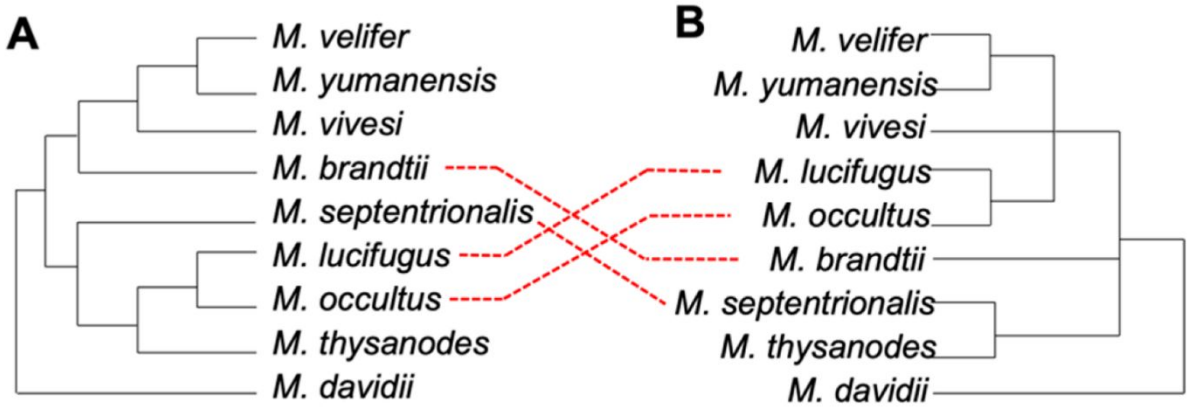
[3] Department of Natural Resource Management and Natural Science Research Laboratory of the Museum of Texas Tech, Texas Tech University, Lubbock, TX 79409, USA

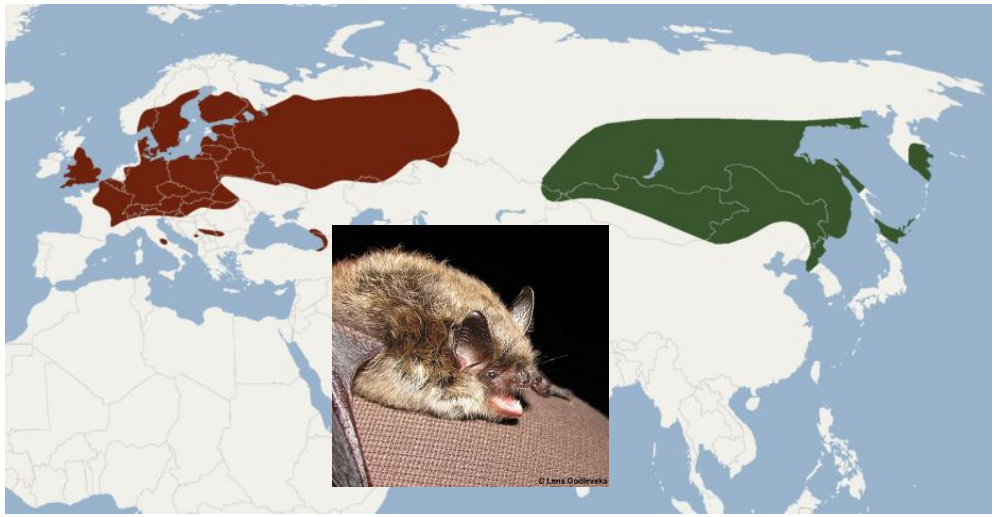
* Author to whom correspondence should be addressed.

Genes **2022**, *13*(3), 399; https://doi.org/10.3390/genes13030399

Myotis phylogenomics

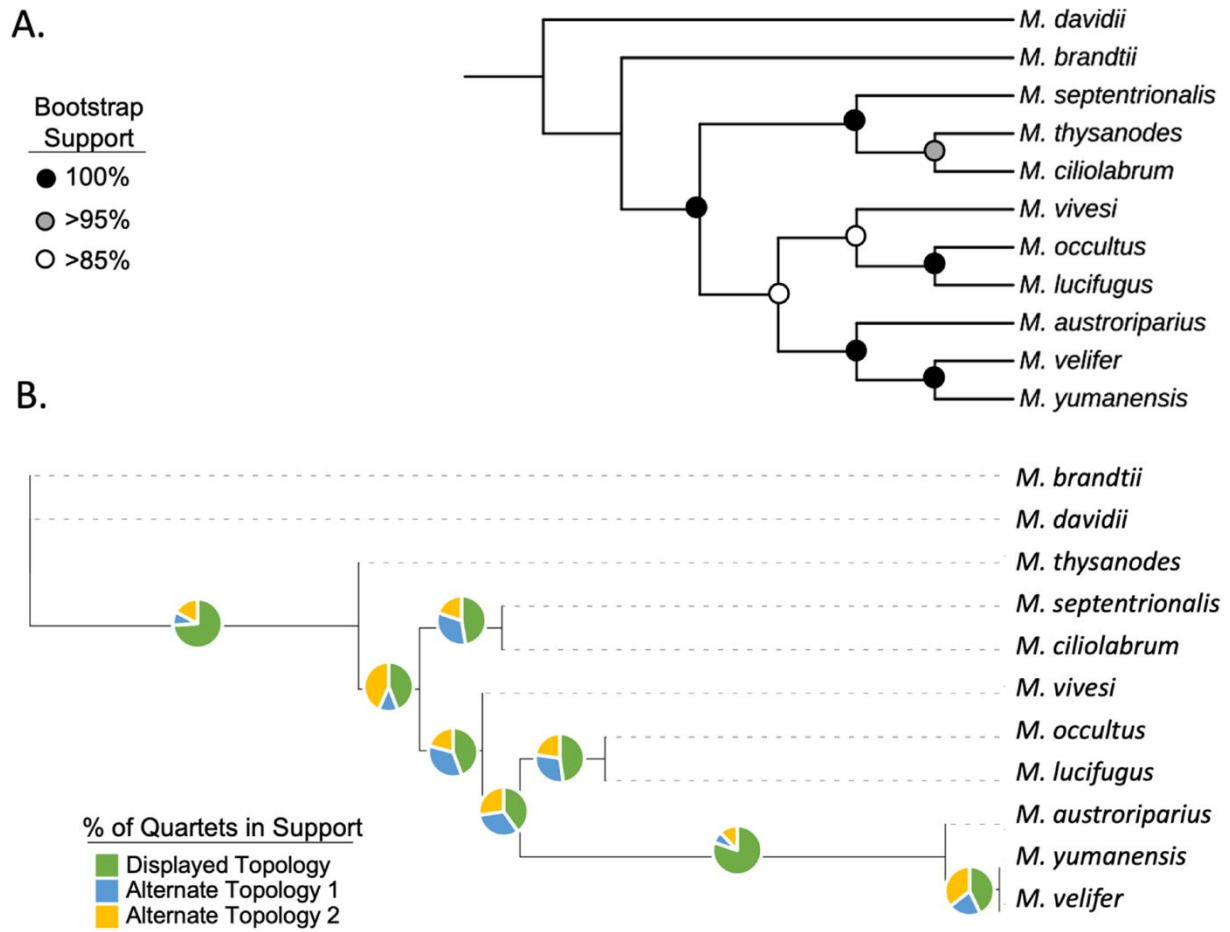
- **Massive recent adaptive radiation**
- **Phylogeny is unclear**





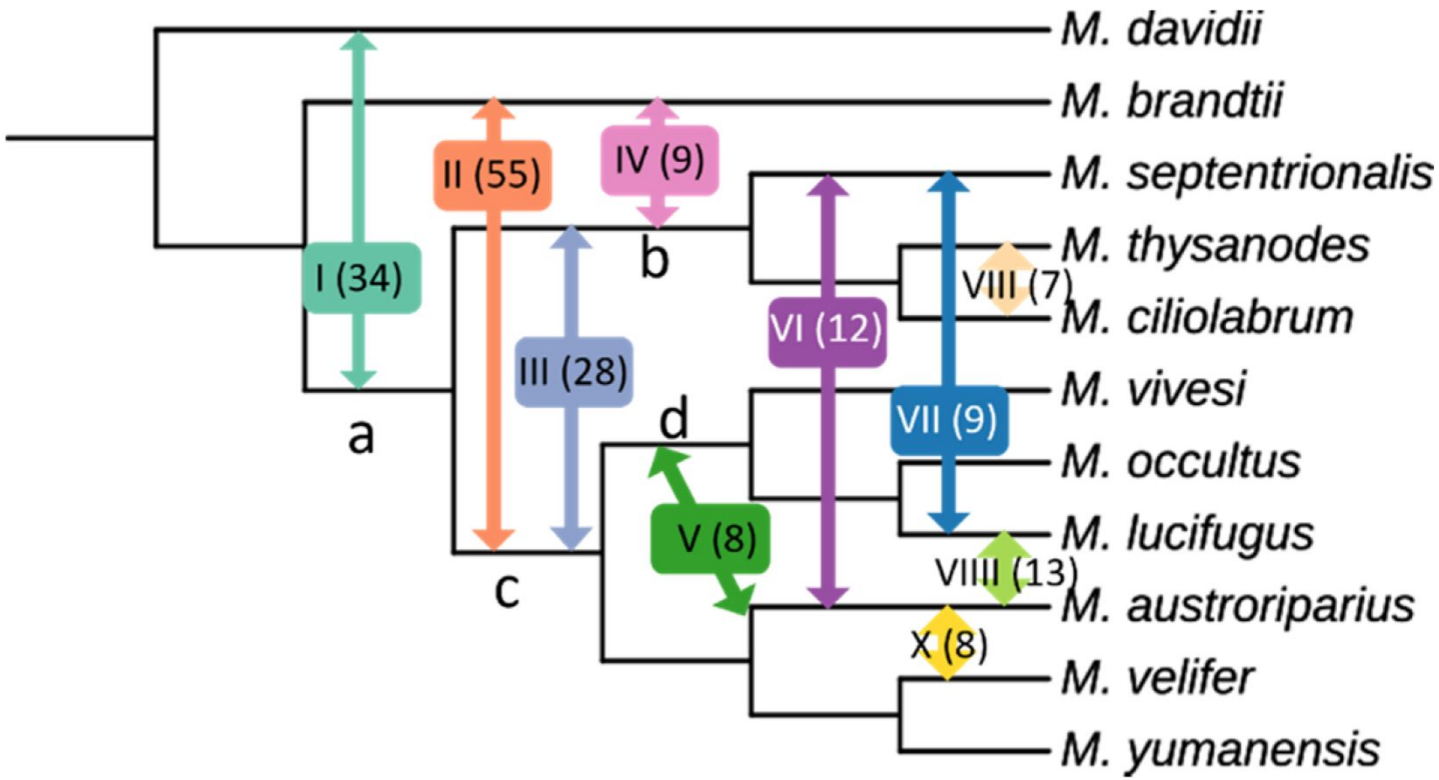
Myotis phylogenomics

- **Applying genome-wide markers allowed us to establish a well-supported phylogenetic hypothesis**



Myotis phylogenomics

- **And identify likely introgression (ancient interbreeding) and/or phylogenetic conflict that help explain the difficulty in understanding the evolution of this group.**



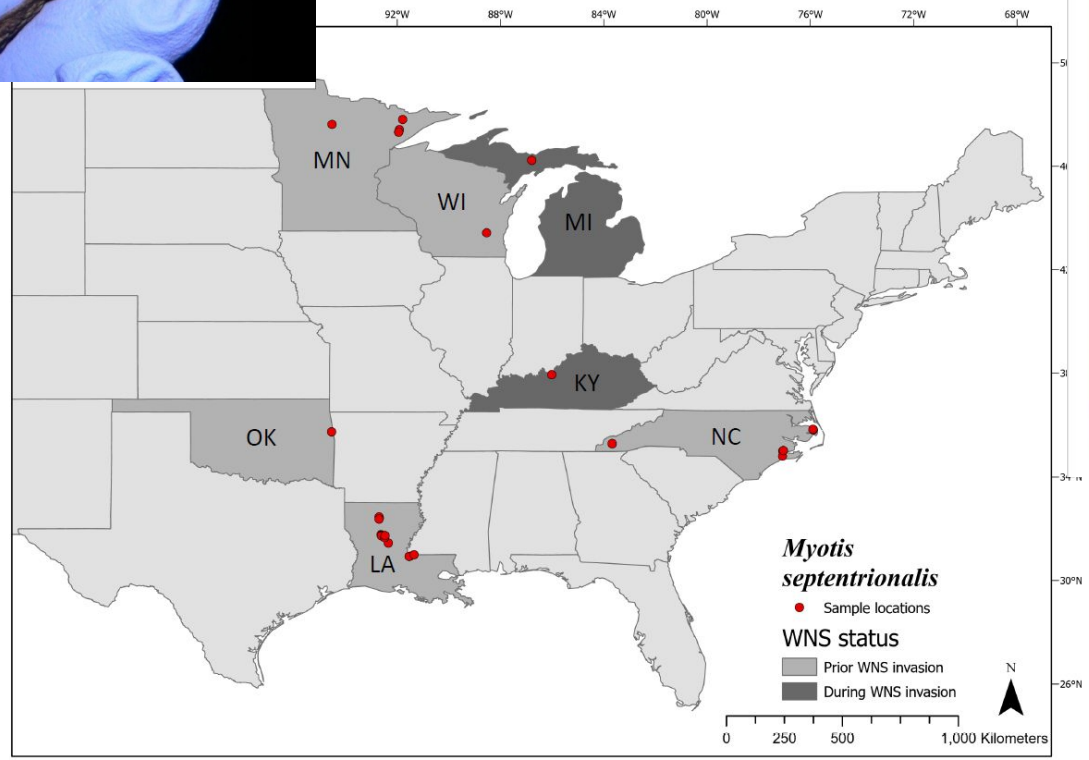
Population health and structure - *Myotis septentrionalis* RAD-Seq

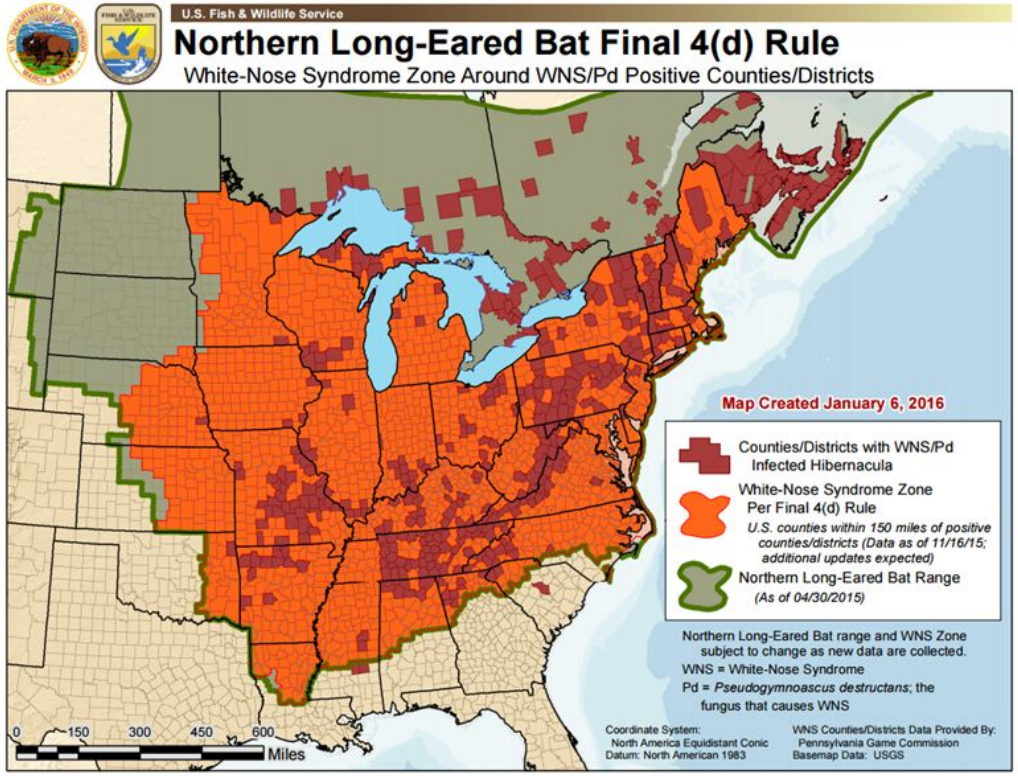
Disentangling genetic diversity of *Myotis septentrionalis*: population structure, demographic history, and effective population size

Jenna R. Grimshaw[1], Deahn Donner[2], Roger Perry[3], W. Mark Ford[4], Alex Silvis[5], Carlos J. Garcia[6], Richard D. Stevens[7,8], and David A. Ray[9*]

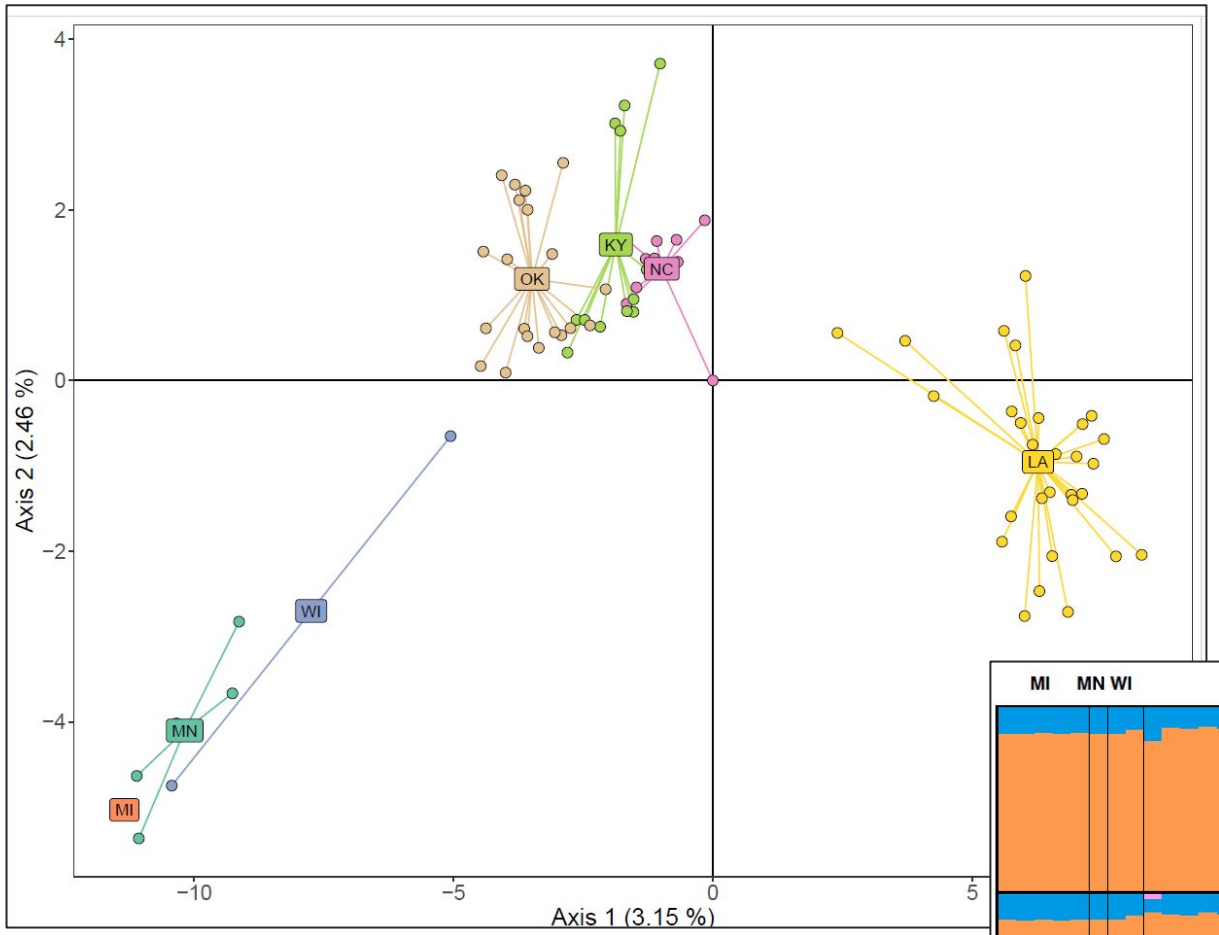
Myotis septentrionalis RAD-Seq

- **Widespread but recently hit hard by white-nose syndrome**

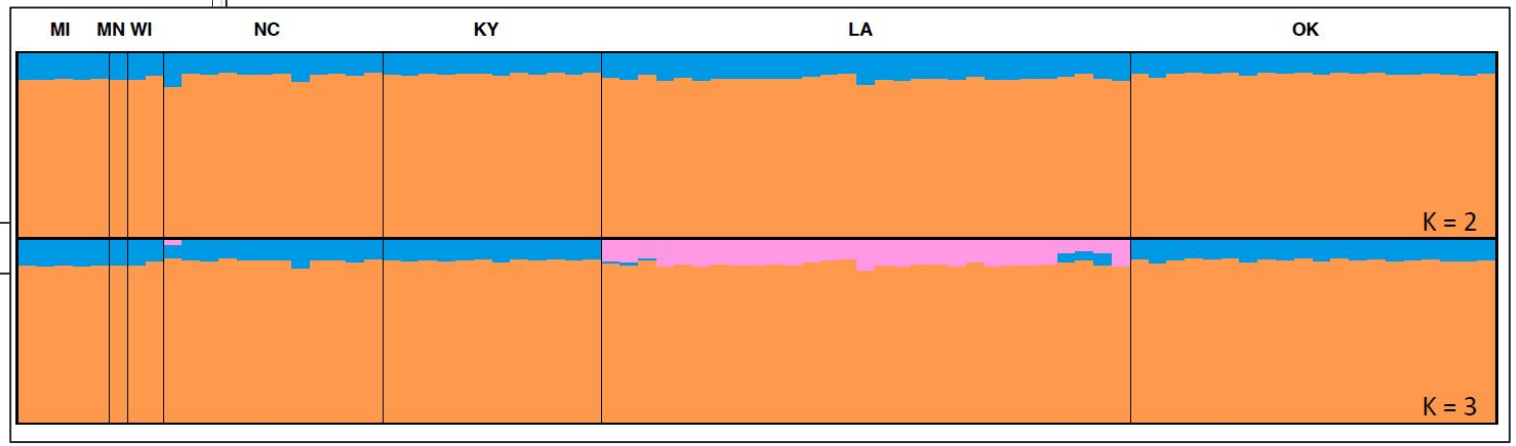




Myotis septentrionalis RAD-Seq



- **Population differentiation is visible, but just barely**
- **LA is exhibiting signs of genetic uniqueness, likely due to isolation from other populations**
- **Suggests panmixia in this species**
- **Data serve as a baseline to investigate impacts of WNS in the future**



The potential is vast

Comparative transcriptomics reveals the evidence of genetic adaptations in the macrotis group (Chiroptera: Rhinolophidae)

Insights into the formation and diversification of a novel chiropteran wing membrane from embryonic development

Comparative transcriptome analysis reveals molecular adaptations underlying distinct immunity and inverted resting posture in bats

Single-cell transcriptome analysis of the *in vivo* response to viral infection in the cave nectar bat *Eonycteris spelaea*

The pale spear-nosed bat: A neuromolecular and transgenic model for vocal learning

LANDSCAPE LEVEL PATTERNS OF GENETIC DIVERSITY OF BAT SPECIES OF GREATEST CONSERVATION NEED IN LOUISIANA

Gene losses in the common vampire bat illuminate molecular adaptations to blood feeding

The evolution of antimicrobial peptides in Chiroptera

Recent expansion and adaptive evolution of the carcinoembryonic antigen family in bats of the Yangochiroptera subgroup

Large-scale genome sampling reveals unique immunity and metabolic adaptations in bats

Comparative genomic analyses illuminate the distinct evolution of megabats within Chiroptera

Genome-wide ultraconserved elements resolve phylogenetic relationships and biogeographic history among Neotropical leaf-nosed bats in the genus *Anoura* (Phyllostomidae)

Phylogenomic Analyses Elucidate the Evolutionary Relationships of Bats

Inferring genetic structure when there is little: population genetics versus genomics of the threatened bat *Miniopterus schreibersii* across Europe

Take home messages

- Genomes are not the answer to every research question but can be powerful tools to help answer many questions
- Genomes assemblies are difficult to obtain but it's getting easier (and cheaper)
- Genomics is bioinformatically intensive
- High-quality DNA is required for high-quality assemblies
- Applications of genomics to biological problems are varied and informative

Questions?

- david.a.ray@ttu.edu
- https://www.davidraylab.com/